Barnes, T., Stamper, J., Madhyastha, T. Comparative Analysis of Concept Derivation Using the Q-matrix Method and Facets. *In J. Beck and T. Barnes (Eds.) Proceedings of the AAAI 21th National Conference on Artificial Intelligence Educational Data Mining Workshop (AAAI2006)*, pp. 21-30. Menlo Park, CA: AAAI Press.

Comparative Analysis of Concept Derivation Using the Q-matrix Method and Facets

Tiffany Barnes, John Stamper

Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC 28223
{tbarnes2, jcstampe}@uncc.edu

Tara Madhyastha

FACET Innovations, LLC
1314 N.E. 43rd St., Suite 207
Seattle, WA 98105
tara@facetinnovations.com

Abstract

We present a pioneering comparison between an expert-driven clustering technique called Facet Theory with the data-driven q-matrix technique for educational data mining. Both facets and q-matrices were created in order to assist instructors with diagnosing and correcting student errors, and each have been used to augment computer-assisted instructional systems with diagnostic information. However, facets are very specific aspects of knowledge, and the decomposition of a topic into facets can be overwhelming to teachers who need this diagnostic help. We present a set of four experiments, demonstrating that the q-matrix educational data mining technique reflects expert-identified conceptual ideas, but does so at a higher level than facets, indicating that a combination of expert-derived and data-derived conceptualizations of student knowledge may be most beneficial.

1. Introduction

Modern educational practice urges teachers to find out what their students think about a topic, prior to instruction. The teacher can then tailor the instruction to the initial conceptions. These techniques are especially recommended for the teaching of science, and for the delivery of computer-aided instruction. The assumption behind this approach is that students think consistently, albeit perhaps erroneously. If this assumption is correct it should be possible to identify objectively defined, consistent patterns in student assessment data. However there are few quantitative methods for mining educational data to identify concepts and the similarities between them.

Intelligent tutoring systems such as those in (Conati, et al., 2002; Heffernan & Koedinger, 2002; Van Lehn & Martin 1998) strive to identify student conceptions, but the majority of these systems require the construction of complex models that are applicable only to a specific tutorial in a specific field, requiring the time of experts to create and then test these models on students. In fact, these are only a few of the tradeoffs ITS system developers face (Murray 1999). One system, REDEEM, was built to ameliorate the time needed to create an ITS, and allow teachers to apply their own teaching strategies in an existing computer-based training (CBT) system, and has been shown to be more effective than a non-expert human tutor in improving student test scores (Ainsworth, et al., 2003). Another system, ASSERT, was built to replace the need for expert model design, using theory refinement to learn student models from behavior, and generate feedback for remediation (Baffes & Mooney, 1996). Similarly, the q-matrix method, as described in (Barnes, 2005) employs knowledge discovery and data mining techniques to automatically assess student knowledge and direct knowledge remediation.

Assessments that reveal student conceptions are often called "diagnostic" assessments to distinguish them from assessments that are designed to identify only a student's level of mastery of a subject. Typically authors of diagnostic assessments have relied upon experts to identify important instructional concepts and to design questions that can identify the common patterns of student thought. One such approach is called "facet theory" (Hunt and Minstrell, 1992). This approach, based on research on student misconceptions, catalogues common observable student ideas within a topic and uses them to create diagnostic questions. Thus, a pattern of student responses to a set of questions should describe to a teacher what conceptions would need to be addressed.

The process of identifying facets from qualitative research is laborious and error prone. The models of student knowledge used to create the lists of misconceptions and knowledge states that students possess are created by experts. The basic idea behind diagnostic instruction is that someone knows the patterns of errors that students have, and their frequency. In almost all fields experts are notoriously inaccurate at estimating the frequency of an event. Unless special precautions are taken unusual or striking ideas stand out, and the commonplace (base rate) is underrated (Fischoff, Licthenstein et al. 1981; Gigerenzer 2000). This may lead to over-diagnosis that will in turn overwhelm a teacher.

For diagnostic assessment to be truly useful, it must be able to scale. Automated techniques for identifying concepts from student data can help identify important ideas that instruction should address. In this paper, we compare the concepts identified automatically from diagnostic question sets using the q-matrix method to the facets they elicit. We find that the q-matrix is able to pull out several important instructional concepts that span many facets, simplifying diagnosis. The mapping from q-matrix concepts to facets is clearest when the question set reliability is high and when students are of higher ability.

The remainder of this paper is organized as follows. Section 1.1 overviews related work. We describe the q-matrix method in depth in Section 2, and facet theory in Section 3. We relate the two methods on the basis of their underlying student models in Section 4. In Section 5, we describe our methodology. Section 6 describes our experimental results. We conclude with final remarks in Section 7.

1.1 Related Work

The q-matrix was devised by Tatsuoka to develop a framework of knowledge structures and states (Birenbaum, et al. 1993; Tatsuoka, 1983), and similar matrices, called skill matrices, have been recently used in intelligent tutoring systems to represent the knowledge components in problems (Koedinger, et al. 2004). In skill or q-matrices, rows are attributes (e.g. tasks, concepts, skills) and columns are items or questions. Each entry in the matrix (1 or 0) indicates whether the attribute is involved in the solution of the item.

Given such a matrix, we can automate identification of a student's knowledge state. However, determination of the attributes contributing to each question requires expert knowledge. In addition, and more importantly, it has been shown that q-matrices constructed by experts do not always accurately reflect patterns of student thought (Hubal, 1992). In other words, we have no evidence that these states and relationships correspond to student understanding.

The q-matrix method is a data-mining algorithm that extracts a q-matrix from student assessment data, to discover "concepts" that influence student behavior, as described in (Barnes, 2005). This is an iterative algorithm that refines and adds concepts by adjusting randomly assigned matrix entries until the total error associated with clustering students by concept states is minimized.

Facet theory, developed by Minstrell (Minstrell 2001) grows out of research on student conceptions. A facet is a small observable piece of knowledge or a strategy that a student uses to make sense of a problem. A tool called DIAGNOSER (Hunt and Minstrell, 1996) is used to diagnose facets that appear in a class of students. Questions may be multiple choice, numerical response, or open-ended. Multiple choice and numerical response ranges are coded to facets automatically.

2. Q-matrix Method

The original inspiration for the q-matrix method came from Tatsuoka et al., who explored student misconceptions in basic math concepts, such as adding fractions (Birenbaum, et al. 1993; Tatsuoka, 1983). The main goal of this research was diagnosis of students' misconceptions, which could be used to guide remediation, assess group performance as a measure of teaching effectiveness, and discover difficult topics (Birenbaum, et al. 1993). Tatsuoka developed a rule space, based on a relatively small set of rules and ideas, in which hypothesized expert rules and actual student errors in fraction addition could be mapped and compared. For example, for -1 + -7, one "rule" is that the student might add the two absolute values. This answer, 8, would them be compared with student answers. This space allowed instructors to map student errors without having to catalog every possible mistake. The expert point in rule space closest to the student response corresponds to the rule the student is assumed to be using. This method improves on other procedural models, by creating a space where all student responses can be compared to expert predictions.

This idea of determining a student's knowledge state from her test question responses inspired the creation of a q-matrix, a binary matrix showing the relationship between test items and latent or underlying attributes, or concepts (Birenbaum, et al., 1993). Students were assigned knowledge states based on their test answers and the constructed q-matrix. An example of a binary q-matrix is given in Table 1. A q-matrix, or "attribute-by item incidence matrix", contains a one if a question is related to the concept, and a zero if not. For example, in this q-matrix, questions q1 and q6 are both related by concept con1, while q1 is also related to q2 and q4 by concept con2. Brewer extended these to values ranging from zero to one, representing a probability that a student will answer a question incorrectly if he does not understand the concept (1996).

Table 1: Example q-matrix

	Questions					
	q1	**q2**	**q3**	**q4**	**q5**	**q6**
con1	1	0	0	0	0	1
con2	1	1	0	1	0	0
con3	1	1	1	0	0	0
con4	1	0	1	0	0	0

Tatsuoka's rule space research showed that it is possible to automate the diagnosis of student knowledge states, based solely on student item-response patterns and the relationship between questions and their concepts. Though promising, the rule space method is very time consuming and topic-specific, and requires expert analysis of questions. The rule space method provides no way to measure or validate that the relationships derived by experts are in fact those used by students, or that different experts will create the same rules.

In 1992, Hubal studied the correspondence between expert-derived q-matrices and student data, and found that these two did not necessarily coincide. In 1996, Brewer created a method to extract a q-matrix from student data, and found that the method could be used to recover knowledge states of simulated students. In (Barnes & Bitzer, 2002), we applied the method to large groups of students, and in (Barnes, 2005; Barnes, Bitzer, & Vouk, 2005) found the method comparable to standard knowledge discovery techniques for grouping student data. In particular, the method outperformed factor analysis in modeling student data and resulted in much more understandable q-matrices, but had higher error than k-means cluster analysis on the data. However, the q-matrix method is preferable to cluster analysis for automated direction of student learning, because human intervention would usually be required to create behaviors to associate with each cluster.

2.1 Q-matrix Algorithm

The q-matrix algorithm, as devised by Brewer in 1996, is a simple hill-climbing algorithm that creates a matrix representing relationships between concepts and questions directly from student response data. The algorithm varies c, the number of concepts, and the values in the q-matrix, minimizing the total error for all students for a given set of n questions. To avoid of local minima, each hill-climbing search is seeded with different random q-matrices and the best of these is kept.

First, c, the number of concepts, is set to one, and a random q-matrix of concepts versus questions is generated with values ranging from zero to one. We then cluster student response data according to "concept states", and compute the total error associated with assigning students to concept states, over all students.

After the error has been computed for a q-matrix each value in the q-matrix is changed by a small amount, and if the overall q-matrix error is improved, the change is saved. This process is repeated for all the values in the q-matrix several times, until the error in the q-matrix is not changing significantly.

After a q-matrix is computed in this fashion, the algorithm is run again with a new random starting point several times, and the q-matrix with minimum error is saved, to avoid falling into a local minimum. It is not guaranteed to be the absolute minimum, but provides and acceptable q-matrix for a given number of concepts.

To determine the best number of concepts to use in the q-matrix, this algorithm is repeated for increasing values of c. The final q-matrix is selected when adding an additional concept does not decrease the overall q-matrix error significantly, and the number of concepts is significantly smaller than the number of questions. This is comparable to the "elbow" criterion in choosing the number of factors for a factor analysis. For this study, q-matrices with an error rate of less than 1 per student were selected. Other built-in criteria could also be used to protect from over-fitting the data.

2.2 Q-matrix Evaluation

In the q-matrix method, student responses are grouped into clusters by concept states. Each cluster in the q-matrix method is represented by its concept state, a vector of bits where the ith bit is 0 if the students do not understand concept i, and a 1 if they do. Each concept state also has associated with it an ideal response vector (IDR). We use the concept state with the q-matrix to determine the IDR. For each question q in the q-matrix we examine the concepts needed to answer that question. If the concept state contains all those needed for q, we set bit q in the IDR to 1, and otherwise to 0. When the q-matrix contains only binary values (not probabilities between 0 and 1), this can be calculated for a concept state c and the q-matrix Q by the following procedure, composing $\neg c$ with Q:

$$IDR = \neg ((\neg c)\, Q)$$

For example, given concept state c = 0110 and the q-matrix Q given in Table 1, $\neg c = 1001$, $(\neg c)Q = 101001$. Therefore, $IDR = \neg((\neg c)Q) = 010110$. This can be explained by viewing $(\neg c)Q$ as all the questions that require knowledge in the concepts that are unknown for a student in concept state c. Thus, the IDR for c is exactly the remaining questions, since none of these require concepts that are unknown for a student in concept state c.

When the q-matrix consists of continuous probabilities, we compute the IDR as explained above, but the negation symbol is interpreted as the probability of the opposite outcome, so in each case where a not appears, we interchange any following values x with 1-x.

A q-matrix is evaluated on its fit to a set of student responses, and is measured as error per student. We now describe the error computation method. First, we create an array whose indices are answer vectors, from 0 up to 2^q-1, where q is the number of questions in the tutorial. We then tally the number of student responses with each answer vector. Then, for each response with at least one student, we compare the response with all IDRs and choose the one closest in Hamming distance. This distance is called the "error" for the student response. We sum all errors over all students to determine the overall error for the q-matrix.

3. Facet Theory

Facet theory, like the q-matrix, was designed to explain student knowledge (Hunt and Minstrell, 1996; Minstrell 2001). A facet is a small observable piece of knowledge or a strategy that a student uses to make sense of a problem. Often students develop intuitive conceptions about a subject, or misconceptions, that interfere with learning. However, many prior conceptions are not technically incorrect, but may be classified as productive or not for learning. Both productive and unproductive ideas are catalogued in "facet clusters", or lists of related facets.

Facets are normally identified and validated through an iterative process. First, content experts survey research on related student misconceptions in the domain of interest at the appropriate age range and develop a set of open-ended questions that will elicit these ideas. Second, student responses to these questions are coded and counted to determine the common (> 10%) facets of response. Third, these responses are loosely organized, using a numbering system, by "how problematic" they are, or how bad it would be to have a student leave the classroom holding those ideas. This is in contrast to the typical classification of student knowledge as correct conceptions versus misconceptions.

A facet beginning with 0 is correct (a "goal facet"), facets in the 40s or 50s usually reflect an incorrect synthesis of some classroom instruction with previous misconceptions, and facets in the 80s and 90s represent very low-level ideas (e.g., an upward slope on a graph means there is a hill).

Hunt and Minstrell developed a WWW tool called DIAGNOSER (Hunt and Minstrell, 1996) to diagnose facets that appear in a class of students. DIAGNOSER question sets contain approximately 7-10 conceptual questions. Most questions are multiple-choice or numerical response questions where choices and numerical response ranges are coded to facets. When a response does not correspond to a facet, either because the multiple-choice distractor was designed that way or because a student enters a numerical value that cannot be matched to some reasoning strategy, the student is asked to repeat the question. Some branching is designed to pinpoint problems when a student exhibits a problematic facet.

The pattern of facets diagnosed for each individual student may be unreliable; typically students alternate between some set of related facets depending on their proficiency in the subject area. Nevertheless, consistency of student reasoning is usually tested at least once by asking the student to select, in a subsequent question, the statement that best corresponds to their reasoning.

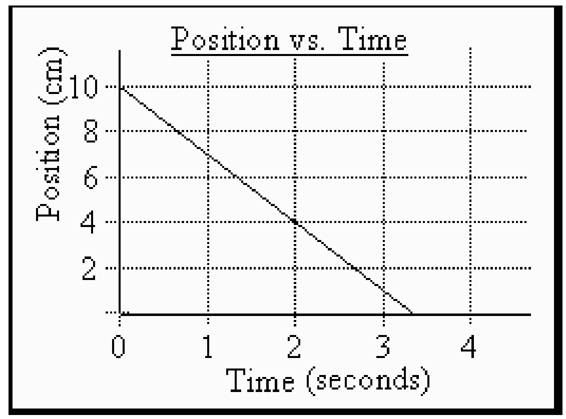
DIAGNOSER is designed to assist teachers in the difficult task of diagnosing and correcting student misconceptions. Teachers often have trouble recognizing the particular misconceptions students might have, and then do not know how to correct them beyond indicating wrong answers. This approach does not usually effect deep conceptual change. On the other hand, teachers can use DIAGNOSER to discover misconceptions and challenge students to explore alternate ideas. Ideally, the teacher will use DIAGNOSER to elicit students' facets of thinking. Then, students will be given a chance to test their ideas with a series of experiments or prescriptive activities. These explorations are designed to challenge students and help them to move towards the target level of understanding. The full suite of DIAGNOSER questions and materials is available at www.diagnoser.com.

Figure 1 shows question Q3 from one of the sets (DS1) that we analyze in Section 6.1. This question is a numerical response question, where possible input values are coded to facets as shown in Table 2. Because it is a position vs. time graph and the object moves 6m in 2 seconds, the correct response is 3.0 m/s. The question is constructed so that students with common misconceptions will give other answers. It is possible that the student can provide an answer that is not coded to a specific facet. Such a facet is "unknown".

Question Q3.
Below is a position versus time graph of the motion of a toy car. What is the speed of the car at t = 2 seconds?

Type your answer in the box below. Your answer must be a number.

_______ meters/second



Answer range classification by Facet:

- **a.** Other **[Unknown]**
- **b.** 3.0-3.0 **[02]**
- **c.** 4.0-4.0 **[71]**
- **d.** 0.0-0.0 **[76]**
- **e.** 2.0-2.0 **[Unknown]**
- **f.** 6.0-6.0 **[71]**

Figure 1. Question Q3 from Determining Speed Set 1

Table 2. Sample "Determining Speed" facets

Facet ID	Description
02	Given position vs. time data, student correctly describes and determines the speed of an object moving uniformly.
71	When asked for the speed at one instant, the student incorrectly reports another quantity or rate. Student reports the position, change in position or distance traveled
76	When asked for the speed at one instant, the student incorrectly reports another quantity or rate. Student reports zero, the object cannot be moving at an instant in time.

4. Comparison of Underlying Student Models

The facet approach is conceptually rather different from the q-matrix approach in its underlying student model. For each question in the DIAGNOSER question set, a facet diagnosis is made. Certain questions are intended to elicit specific facets. Each answer a student may select relates to only one facet. In the q-matrix approach, each question is related to a set of concepts. However, in the q-matrix approach the answer is rated only as right or wrong, and is seen to have resulted from a combination of knowledge of each concept related to that question. A q-matrix characterization of student knowledge comes from a profile of all concepts across all the questions, while facets are determined for each individual answer. The relationship of these two approaches relies upon how consistently students respond.

If students reason coherently across different questions (e.g., if they read every graph as a map of motion) the q-matrix method would reflect these consistent thought patterns, and we hypothesize that the resultant concepts would be comparable to facets. However, if students do not reason coherently, or if they respond according to patterns that the questions are not designed to measure (e.g., suppose a student considers axes labels only half the time, or answers every question with "c"), q-matrix error would increase, and the extracted concepts would be less clearly identifiable as meaningful knowledge patterns or facets.

According to the facet approach, students exhibit one of a small subset of related facets at any given time, and will move between them on similar questions. However, evidence suggests that students of higher math ability reason more coherently across questions, even when their responses are incorrect (Madhyastha et al, 2006). This suggests that the q-matrix method would work best, in the sense of mining pedagogically useful concepts, for question sets taken by students of higher ability.

5. Procedure

We apply the q-matrix method and facet theory analysis to data collected from three physics question sets from the DIAGNOSER system from January 2004-May 2005. These sets were taken by students and some adults in all grade levels. Our hypotheses are: 1) the extracted q-matrices and facet theory analysis will show overlap in the concepts derived, and 2) the concepts derived from the q-matrix method will be clearer as question sets become more reliable and as students respond more coherently.

The question sets we used are: Determining Speed Set 1 (DS1), Determining Speed Set 2 (DS2), and an extended set called New Determining Speed Set 1 (NDS). All of these sets are designed for diagnosing student ideas surrounding speed for grades 6-12, and are aligned with national standards for physics. Target concepts include being able to describe and calculate the speed of an object from: position vs. time graphs, speed vs. time graphs, data tables, and from dots drawn by a moving car at uniform intervals on a strip of paper (a "dot car" representation).

Determining Speed Set 1 (DS1) has 9 questions and Determining Speed Set 2 (DS2) has 10 questions. This length is typical of DIAGNOSER question sets. However, because these sets are designed so that students who have mastered the concepts obtain near perfect scores, they provide little reliable information about high achieving students. For this reason, the New Determining Speed Set 1 was developed as an experimental set for use with other psychometric models. This set has 9 question "bundles", most of which consist of a base question from Determining Speed Set 1, a follow-up challenge question for students who answer the base question correctly, and a follow-up repeat question, extremely similar to the original, for students who answer the base question with an unknown facet. This set is designed to be more reliable than the others. This new set and Determining Speed Set 2 may be viewed at www.diagnoser.com.

We compare facets with extracted q-matrices for 5 datasets: DS1, DS2, NDS, and DS1 separated into grade levels 7-8 and 11-12. In all data sets, conditional questions with few responses were eliminated from the data set. The only conditional question in DS1 is question 5, and this question was eliminated from DS1 for all experiments. NDS contains several conditional questions, and questions are numbered as 1, 1r, and 1c, for example, to indicate that 1r and 1c are repeat and challenge questions for question 1. Missing values in these bundles were determined by applying the following assumptions: 1) if question 1 is correct, the next question given is 1c, and 1r, although not administered, is assumed to be correct, and 2) if question 1 is incorrect, the next question given is 1r, and 1c is assumed to be incorrect.

There were no missing values for questions administered to all students in a set, since students are not allowed to advance to the next question without answering the current one. Data for each set consists of a list of student responses already coded into facets. We converted these facets into bits representing whether the student answer was correct or incorrect. For goal facets numbered 01-03, these answers were correct, and coded as ones, while all other problematic facets were converted to incorrect answers, or zeroes. Only the first question attempt for each student (in the case of repeated identical questions, which occur only in DS1 and DS2) was considered.

6. Results and Discussion

We present the results of extracting q-matrices for students studying Determining Speed Sets 1 & 2 (DS1 & DS2) from January 2004-May 2005, and New Determining Speed Set 1 (NDS) from November 2005-March 2006. We also created q-matrices for students at two grade levels (7-8 and 11-12) taking the DS1 set between January 2004-January 2006. In each section below, we compare the extracted q-matrices with the diagnosed facets.

6.1. Determining Speed Set 1 (DS1)

We applied the q-matrix method to DS1, including responses from 1502 students. Question Q5 was conditional, and few students took it, so it was dropped from the analysis. For the remaining 8 questions, a five concept solution resulted in the lowest error, about 1 per student. Table 3 shows the resulting q-matrix with concepts labeled A-E. Table 4 shows the intended facets covered by each question of the set. The facet labeled -1 is unknown, meaning that the cluster of students this facet describes is using a strategy that experts could not identify.

Table 3. DS1 Q-matrix, 1502 students, error 1/student

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9
A	0	0	1	0	0	0	1	1
B	1	1	1	0	1	0	0	0
C	0	0	1	0	1	1	0	0
D	0	0	1	1	0	0	0	0
E	0	1	0	0	1	0	0	1

Table 4. Facet map for Determining Speed Set 1

Facet	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
-1	1	1	1				1		
01								1	1
02		1	1						
03	1			1	1	1	1		
41							1		
42				1					
50							1		
51				1					
52						1			
70						1			
71			1						
72								1	1
73								1	1
74								1	1
75								1	1
76			1						
81		1							
82	1								
83					1				
84	1			1					
90		1		1					

The first thing to note is that question Q3 appears in all but one of the identified concepts. Based on this q-matrix, we predicted that this question was the most difficult in the set. Question Q3 is a numerical response question that asks students to calculate the speed of an object at an instant using a position vs. time graph. This is the most difficult question in terms of aggregate response: only 27% of students answered it correctly. As indicated by the fact that it relates to three concepts, question 4, which asks students to describe the trip of an object from a speed vs. time graph, is the second most difficult (53% of students answered it correctly). Between 60-77% of students answered each of the remaining questions (related to 1-2 concepts) correctly. It may be that question 9 is essentially acting as a "consistency" filter, cutting off students who are really answering with piecemeal knowledge? In other research we find that students of higher ability answer more systematically.

When comparing the concepts in Table 3 to the facets in Table 4, we must remember that the goal facets 01-03 are all correct and thus should not be reflected as q-matrix concepts. There is no obvious correspondence from facets to the q-matrix except for concept A, which maps directly to the union of facets 71-76. In DS1, questions Q7 and Q8 are paired, where question Q8 asks students for reasoning to explain their answer to Q7. Concept A corresponds to the ability to give the speed at an instant in time.

Since the relationship between facets in concepts B-E is not as clear, we interpreted these concepts by examining the percentage of students with each problematic facet and noting the skills required by the students to answer the questions. Concept B is possessed by students who know how to interpret a position vs. time graph to describe speed for straight and sloped segments, can calculate the speed from a position vs. time graph, and can describe speed from a dot car representation of motion. The problematic facets that occur most frequently (70, 71, 81 and 82) are, in context, all related to confusion of position and speed graphs. This is a very common problem that students have when learning this material. We believe that Concept B corresponds to not confusing position vs. time graphs and speed vs. time graphs or other representations.

Concept C is demonstrated by proficiency on questions Q3, Q6 and Q7. These ask students to describe the motion of a car with constantly spaced dots, look at a table of speed data and describe motion, and look at a position/time graph and describe the speed at an instant. The most frequently occurring problematic facets in these questions are 70 and 71. There is clearly overlap with concept B here – both include Q6 and Q3. Question Q7 is unique in that it asks about a data table. Concept C corresponds to the ability to interpret three types of speed/time information.

Concept D includes questions Q3 and Q4. Question Q4 asks if students can accurately describe motion over a whole trip. Combined with the most difficult question (Q3), concept D corresponds to being able to describe motion over a trip.

Finally, concept E is interesting in that it includes questions Q2, Q6 and Q9. Since these questions do not seem to relate to a common set of facets, we conclude that concept E seems to be the "good test-taking" concept. In other words, mastery of concept E is shown by students who are good at test taking, while missing Q2, Q6, and Q9 together might indicate students who are attaining answers by guessing. We note that, in generating any types of clusters, there is often one cluster that groups some diverse elements that don't quite fit with other clusters. This may also be a plausible explanation of concept E.

6.2. Determining Speed Set 2 (DS2)

We next applied the q-matrix method to DS2, including responses from 335 students, resulting in a q-matrix using 4 concepts for the 10 questions and attaining an error rate of 1.28 per student. Table 5 shows the resulting q-matrix with concepts labeled A-D. Table 6 shows the intended facets covered by each question of the set.

Table 5. DS2 Q-matrix, 335 students, error: 1.28/student

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
A	0	0	0	0	1	0	1	1	1	0
B	0	0	0	0	0	0	0	1	0	1
C	1	1	0	0	0	0	1	1	1	0
D	0	0	0	1	1	1	1	1	0	0

Table 6. Facet map for Determining Speed Set 2

Facet	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
-1		1								1
01	1	1								
02								1	1	
03			1	1	1	1	1			1
41			1							
42			1					1		
51	1		1		1	1	1			
52			1							
70									1	1
71								1		
72								1	1	1
73	1	1								
74		1		1						
75	1	1								
76								1		
80	1	1						1		
81							1			
82									1	
83					1	1				
84				1						
90						1				
91					1		1			
92				1						

Question Q8 occurs in each concept. It is orthogonal to question Q3 in DS1, and asks students to calculate an instantaneous speed from a position vs. time graph. Only 13.8% of students answer it correctly. Question 7, which asks students to describe the motion of a car during a portion of a position vs. time graph depicting its trip, is the second hardest (52.5% of students answer correctly), and this is reflected by its relation to three concepts in the q-matrix. The most frequent error is to say that the car is speeding up. As with DS1, this is a frequent problem.

Concept A covers all the questions that have to do with reading the speed off of a position vs. time graph, except for Question 5. Its grouping with concept A may be simply due to its difficulty level – all the questions in this concept involve two or more concepts, meaning that they are the most difficult in this set. The meaning concept A in DS2 is similar to that of concept A identified in DS1. Concept B is extremely interesting in that although it does not map cleanly on to the facets, it includes the most difficult question (Q8) in addition to the only question that asks about the dot car representation (Q10). Although the facets "covered" by the dot-car question overlap with others, this indicates that this may be a different concept. Students might not have been exposed to this representation in their studies. Concept C describes be students who do not confuse position and time graphs. Together, concepts B and C from DS2 map on to concept B derived from DS1. Like concept D in DS1, Concept D in DS2 includes students who can describe motion from a speed vs. time graph, except for Question 8.

We find this correspondence between concepts extracted for DS1 and DS2 reassuring in assessing the performance of the q-matrix method, since both of these question sets are addressing similar pedagogical material.

6.4. Determining Speed Set 1-Grades 7-8 & 11-12

One assumption behind the q-matrix method is that students reason fairly consistently across questions, so that if they have the concepts necessary to answer a question, they are likely to do so correctly. However, this is one model of student reasoning that may not fit all students. In particular, there may be a relationship between ability and the probability of a student reasoning consistently. If this is true, we would expect the q-matrix method to derive concepts that are closer to the expert-defined facets when mining data from students of generally higher ability. In other words, the concepts mined by the q-matrix should correspond more closely to the intended facet diagnosis as students answer questions more consistently. We have found that consistency of student response increases with student ability (Madhyastha et al, 2006). To test this hypothesis, we examined subsets of students at the higher and lower end of the grade range for DS1. We considered separately students in grades 7-8 and grades 11-12. Note that Table 4 lists the intended facets covered by each question of DS1, for both data sets.

The data for students in grades 7-8 includes responses from 468 students, resulting in a q-matrix using 4 concepts for the 8 questions and attaining an error rate of 0.9 per student. Table 7 shows the resulting q-matrix. Labels A,B, and E in Table 7 correspond to those same concepts extracted for DS1, and concept C2D2 is so labeled since it combines concepts C2 and D2 found in the 5-concept q-matrix for G11-12, as given in Table 9.

Table 7. G7-8 Q-matrix, 4 concepts, error: 0.9/student

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9
A1	0	0	1	0	0	0	1	1
B	1	1	1	0	1	0	0	0
C2D2	1	0	1	0	1	1	0	0
E	0	0	1	1	0	0	0	0

We also extracted q-matrices for DS1 for 1035 students in grades 11-12. Table 8 lists the 4-concept q-matrix for this set, which has an error rate of 1 per student. Table 9 shows the extracted q-matrix for 5 concepts, and a total error of 0.9 per student. Using the criteria of about 1 error per student, either of these q-matrices would be acceptable. We list both here to compare G7-8 with G11-12.

Table 8. G11-12 Q-matrix, 4 concepts, error: 1/student

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9
A1	0	0	0	0	0	0	1	1
B	1	1	1	0	1	0	0	0
C1	0	0	1	1	1	0	0	0
E	0	1	1	0	0	1	0	0

Table 9. G11-12 Q-matrix, 5 concepts, error: 0.9/student

	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9
A1	0	0	0	0	0	0	1	1
B2	1	1	1	0	0	0	0	0
C2	0	0	1	0	1	0	0	0
D2	1	0	1	0	0	1	0	0
E2	0	0	1	1	0	0	0	0

In all three q-matrices, labels B, and E correspond to concepts B, and E for DS1, while concepts A1, C1, and B2-E2 are close to those with the same letters in DS1. Although the concepts derived are similar for both groups G7-8 and G11-12, they are far more refined in the upper level grades. Note that, for grades 7-8, question Q3 is still stumping most students (as evidenced by its relation to all 4 extracted concepts), while for grades 11-12, concept A1 is not related to question Q3, as concept A is for G7-8 and DS1. This indicates more sophistication for higher-level students, and the resulting concept A1 is a purer representation of knowledge, corresponding to Facets 01, and 72-75, which are goal and problematic facets for determining the speed at a particular instant. The difference between concepts A and A1 is that students with concept A1 no longer confuse position and speed, or these students are more purely described using concept B2.

Concept B (the ability to distinguish between position and speed) is a very important one. The 5-concept q-matrix for the upper grades separates B out from interpreting the dot car representation (concept C2) from position and speed, resulting in concept B2. In the lower grades, the best fit includes fewer concepts, and many are commingled. Concept D2 combines describing motion from a position and speed graph and from a speed/time table with the most difficult question, Q3.

Although the more concepts used to describe students, the more refined these concepts will be, we find it significant that the refined (5 concept) q-matrix for G11-12 is separating the dot-car representation from position versus speed, and both the 4- and 5- concept q-matrices for upper level students separate speed determination from confusing position and speed (concept A1). This supports our hypothesis that q-matrices derived from more consistent student responses more clearly delineate facets.

6.3. New Determining Speed Set 1 (NDS)

We next applied the q-matrix method to NDS questions 1-8p, including responses from 830 students, resulting in a q-matrix using 4 concepts for the 16 questions and attaining an error rate of 2.4 per student (but still around 15%), given in Table 10. This is not the optimal q-matrix but we found that q-matrices with more concepts were tending to overfit the data, due to the difficulty of questions 2c, 3, 3r, and 5c, which were answered correctly by only 56 students. Refer to Table 4 for the facets for each question.

We summarize the goals of each bundle, which correspond to questions 1-8 in DS1 as follows:

1. Describe the motion of an object from a position vs. time graph
2. Describe the speed of an object from a portion of a position vs. time graph
3. Identify speed from a position vs. time graph
4. Describe motion from a speed vs. time graph
5. Describe motion for part of the journey of an object on a speed vs. time graph
6. Describe motion from a dot car trace (a car that leaves dots on a strip of paper at regular time intervals)
7. Describe motion from a table
8. Identify speed from a speed vs. time graph

We note that in Table 7, each concept relates to questions and repeats the same way (e.g. 1 and 1r are both 0's or both 1's for all concepts). We also note that, outside of questions 2c, 3, 3r, and 5c, which are the most difficult, most questions relate to only one concept, and three of these to B3, which highlights facets 01, and 72-76, and is a crucial concept dealing with relating Exceptions to this are question 2, 2r, and 6, which all also relate to C3, which covers describing motion from graphs (bundles 4-6) and describing speed from a position versus time graph. Concept D3 involves describing motion over a whole trip, and also groups several of the most difficult questions together. Concept A3, corresponds to concept DS1-A, the ability to determine speed at an instant in time. We note that question 7 relates to no extracted concepts, indicating that it was answered correctly by almost all students in the data set.

Table 10. NDS Q-matrix, 4 concepts, error: 2.4/student

	1	1r	1c	2	2r	2c	3	3r	4	4c	5	5c	6	7	8	8p
A3	0	0	0	0	0	1	1	1	0	0	0	1	1	0	1	1
B3	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0
C3	0	0	0	1	1	1	1	1	1	1	0	1	1	0	0	0
D3	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0

We verify that the q-matrix extracted for NDS does not differ greatly from that for DS1 in Table 11, where we have omitted question Q5 and all repeat and challenge questions from NDS. Concept A3 differs in just one position (Q6) from DS1-A. Concept B3, like concept B2 above, removes dot-car representations from DS1's concept B. C3 does not correspond well to DS1, which may indicate some fitting to the other questions. Concept D3 is identical to concept DS1-D.

Table 11. NDS Q-matrix omitting Q5, repeats & challenges

	Q1	Q2	Q3	Q4	Q6	Q7	Q8
A3	0	0	1	0	1	0	1
B3	1	1	1	0	0	0	0
C3	1	1	1	1	0	0	0
D3	0	0	1	1	0	0	0

The q-matrices resulting from the analysis of NDS were not as easy to interpret as we expected. We hypothesized that this data set would be more reliable and reflect facets more clearly than DS1. Through this analysis, however, we have determined that students are not performing as well as we may have hoped on the challenge questions and this performance may be affecting their performance on other items. NDS is a newer set of questions, and the interface for transferring between questions is not as smooth as that for DS1. We also note that students may have gotten into a rhythm of problem-solving in DS1 that may have been disrupted by our conditional branching in NDS, and that this disruption, along with the interface differences, may have affected student performance.

6.5 Q-matrix Method Results

This study serves as evidence of the robustness of the q-matrix method with respect to class size, student grade level, and topic. In our previous work (e.g. Barnes, 2005), we successfully applied the q-matrix method to data collected from 200 college-level discrete mathematics students, but hypothesized that the method could apply in a general topic setting and would scale to larger data sets. We have effectively applied the q-matrix method to question sets administered to groups of 300-1500+ physics students from 7^{th} grade and up. Although the running time of the method is approximately S*Q*C, where S=students, Q=questions, and C=concepts, even with low Q values (8-16), running time becomes long with more than 8 concepts with our Java implementation on a standard modern PC. We plan to address this through improved initialization options, such as initializing the q-matrix with student data values, and through implementing the algorithm in C.

The DIAGNOSER question sets are designed with 9-10 main questions and some repetitive and challenge items, while in previous work we have applied q-matrices to sets of 5-10 questions. It is still an open question whether the q-matrix method will prove effective in analyzing data from much larger question sets. We suspect that the improvements we plan for the method will be essential in ensuring the convergence of the method to acceptable solutions with large question sets.

7. Final Remarks

This study provides one of the first comparisons of an extensive expert decomposition of student knowledge with an automated extraction method. Our results suggest that automated extraction can be interpreted by area experts and provide useful complementary information to an expert-derived system.

Barnes (2005) and Hubal (1992) found that the q-matrices elicited from experts after question sets were created do not necessarily relate to student performance. However, with the data sets created for DIAGNOSER, we were able to find some correspondence between existing expert decompositions and extracted q-matrices.

Q-matrices may even explain higher-level concepts than facets, and in each experiment, q-matrix concepts are reflecting the most important concepts in learning to determine speed from different representations.

Facets are intentionally very low-level constructs. Nevertheless, the breadth of facet patterns diagnosed in the classroom must be simplified to be of use to a teacher. By identifying the relationships between diagnostic questions, q-matrices may explain higher-level concepts than facets. In each experiment, q-matrix concepts are reflecting the most important patterns of error that students typically bring with them to the classroom. We suggest that augmenting DIAGNOSER's feedback with q-matrix concepts may add an important dimension to teacher feedback and may make understanding student behavior easier for classroom teachers.

One problem with using facets for diagnosis is that students may not reason consistently across questions, resulting in diagnoses that are fleeting. Q-matrix concepts help to isolate the patterns of response that occur more consistently. As we found by examining q-matrices resulting from mining students of higher and lower grade levels, the concepts identified from the higher grade level students are more distinct versions of the ones found from the group as a whole.

References

Ainsworth, S.E., Major, N., Grimshaw, S.K., Hayes, M., Underwood, J.D., Williams, B. & Wood, D.J. 2003. REDEEM: Simple intelligent tutoring systems from usable tools, in T. Murray, S. Blessing & S.E. Ainsworth (eds.). *Advanced Tools for Advanced Technology Learning Environments*. pp. 205-232. Amsterdam: Kluwer Academic Publishers.

Baffes, P. & Mooney, R.J. 1996. A novel application of theory refinement to student modeling. *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (AAAI-96), pp. 403-408, Portland, OR, August 1996.

Barnes, T. 2005. The q-matrix method: Mining student response data for knowledge. In *Educational Data Mining: Papers from the 2005 AAAI Workshop*, ed. Joseph E. Beck, pp. 39-46. Technical Report WS-05-02. American Association for Artificial Intelligence, Menlo Park, California.

Barnes, T. & Bitzer, D. 2002. Fault tolerant teaching and automated knowledge assessment. *Proceedings of the ACM Southeast Conference (ACMSE 2002)*, pp. 125-132, Raleigh NC, April 2002.

Barnes, T., D. Bitzer, & M. Vouk. 2005. Experimental analysis of the q-matrix method in knowledge discovery. *International Symposium on Methodologies for Intelligent Systems*, pp. 603-611, Saratoga Springs, NY, May 2005.

Birenbaum, M., Kelly, A., & Tatsuoka, K. 1993. Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education*, 24(5), 442-459.

Brown, J.S. & Burton, R. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2, 155-191.

Conati, C., Gertner, A. & VanLehn, K.. 2002. Using Bayesian networks to manage uncertainty in student modeling. *User Model. User-Adapt. Interact*. 12(4): 371-417.

Fischoff, B., S. Licthenstein, et al. 1981. *Acceptable Risk*. Cambridge; New York, Cambridge University Press.

Gigerenzer, G. 2000. *Adaptive Thinking: Rationality in the real world*. Oxford, Oxford U. Press.

Heffernan, N.T. & Koedinger, K.R. 2002. An intelligent tutoring system incorporating a model of an experienced human tutor. *Intelligent Tutoring Systems,* p. 596-608.

Hubal, R. 1992. Retro-adaptive testing and evaluation system. Computer Science Masters Thesis, North Carolina State University.

Hunt, E. & Minstrell, J. 1996. Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education*, 2, pp. 123-162.

Koedinger, K. R., Aleven, V., Heffernan. T., McLaren, B. & Hockenberry, M. 2004. Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. *Proceedings of 7th Annual Intelligent Tutoring Systems Conference*, Maceio, Brazil. pp. 162-173.

Madhyastha, T., Hunt, E., Kraus, P., & Minstrell, J. 2006. The relationship of coherence of thought and conceptual change to ability. *2006 American Educational Research Association Annual Meeting*, San Francisco, CA, April 2006.

Minstrell, J. 2001. Facets of students' thinking: Designing to cross the gap from research to standards-based practice. *Designing for Science: Implications for Professional, Instructional, and Everyday Science*. K. Crowley, C. D. Schunn and T. Okada. Mahwah, Lawrence Erlbaum Associates.

Murray, Tom. 1999. Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, 10: 98-129.

NovaNET educational network. Online. http://www.pearsondigital.com/novanet/

Sellers, J. 1998. An empirical evaluation of a fault-tolerant approach to computer-assisted teaching of binary relations. Computer Science Masters Thesis, North Carolina State University.

Tatsuoka, K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.

VanLehn, K., & Martin, J. 1998. Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence and Education*, 8(2), 179-221.