Data Mining Patterns of Thought

Earl Hunt*   Tara Madhyastha†

*Department of Psychology, University of Washington, Seattle WA
†FACET Innovations, Seattle WA
ehunt@u.washington.edu, tara@facetinnovations.com

Abstract

Modern educational and psychological measurements are governed by models that do not allow for identification of patterns of student thought. However, in many situations, including diagnostic assessment, it is more important to understand student thought than to score it. We propose using entropy-based clustering to group responses to both a standard achievement test and a test specifically designed to reveal different facets of student thinking. We show that this approach is able to identify patterns of thought in these domains, although there are limitations to what information can be obtained from multiple choice responses alone.

.

1. Introduction

Psychometric evaluations are widely used in a variety of areas, particularly to measure student achievement. However, these evaluations assume that a student approaches a problem or a test in essentially the same way. For example, the underlying model of item response theory assumes that the probability of getting an item correct is determined by a skill level unique to the individual and an item difficulty unique to the item; this is how performance on different tests among different populations can be compared. Items that do not fit the model are necessarily dropped through a "Darwinian" process of test creation. In the area of diagnostic assessment, where the goal is not merely to place a person along one or more axes of achievement but rather to provide a specific educational intervention to help them to progress, it is more important to identify student conceptualizations than to grade them.

The remainder of this paper is organized as follows. Section 2 outlines existing models and related work. In Section 3 we introduce DIAGNOSER, an example of a diagnostic learning environment, which illustrates the need for identifying patterns of thought in diagnostic assessment. Our approach, described in Section 4, uses entropy-based clustering to identify patterns in student answers to multiple-choice questions. In Section 5 we apply this clustering algorithm to a standardized science achievement test (the Washington State Assessment of Learning – WASL). In Section 6, we cluster responses on a diagnostic assessment test designed to elicit different reasoning patterns. We conclude with directions for future work in Section 7.

2. Motivation and Related Work

Modern educational and psychological measurements are dominated by two mathematical models, Factor Analysis (FA) and Item Response Theory (IRT). FA operates at the level of a test, i.e., a collection of questions. The basic assumption of FA is that the test score of individual *i* on test *j* is determined by

$$x_{ij} = \sum_{k=1}^{K} w_{kj} f_{ik} + e_{ij}$$

where the f_{ik} terms represent the extent to which individual *i* has underlying ability *k*, and the w_{kj} terms represent the extent to which the ability *k* is required for test *j*. The e_{ij} term is a residual to be minimized. The relative pattern abilities required for the test, i.e. the $\{w_{kj}\}$, is constant across individuals. This amounts to an assumption that all individuals deploy their abilities in the same way on each test. Put another way, we can imagine a test taking ability for individual *i* on test *j*

$$B_i(j) = \sum_{k=1}^{K} w_{\mathrm{jk}} f_{ij}$$

that characterizes how well person *i* is prepared for test *j*. Note that the *f* here is not the same *f* variable referred to in the first equation.

IRT operates at the item level within a test. Consider the j^{th} item on a test. This item is assumed to have a characteristic difficulty level, θ_j. The i^{th} examinee is assumed to have skill level B_i on the same scale. (The equation above would be used to connect skill level to factor analysis.) In the basic three parameter IRT model the probability that person *i* will get item *j* correct is

$$P(\theta_j,B_{i,}) = c_j + (1\text{-}c_j)\frac{e^{Da_j(B_i-\theta_j)}}{1+e^{Da_j(B_i-\theta_j)}}$$

where D is a scaling factor, a_j is an item discrimination parameter and c_j is a “correction for guessing parameter”. A consequence of this model is that the relative order of difficulty for any pair of items on a test must be the same for all individuals.

Clearly these models do not allow for idiosyncratic patterns of thought, where different people attack problems in different ways. Nevertheless, the models fit most large educational and psychological tests. Why? Test makers regard tests and items that do not fit these models as “bad” tests or items; thus, tests that fit these models cannot reveal patterns of thought.

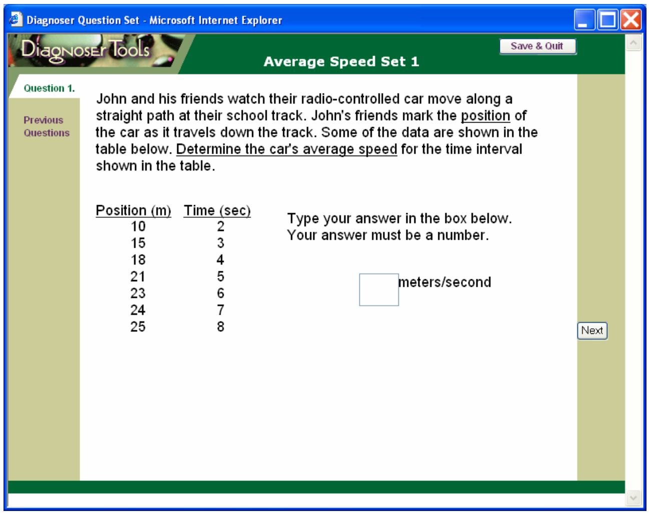
Recently, Bayesian approaches have been used to assign people to patterns to produce more diagnostic assessment information [8, 9]. These have in common a set of hypothesis about student reasoning, and probabilistically assign students to one of a fixed set of strategies. They work very well when the strategies, which must be determined through other methods, are known. However, they fail when a student does not fit the underlying model.

3. DIAGNOSER: A Diagnostic Learning Environment

The DIAGNOSER system (Hunt and Minstrell, 1996) is a world wide web (WWW) based system designed to be used by a teacher to diagnose student difficulties in science. The system consists of short sets of questions designed to elicit middle-school and high-school student thinking around specific concepts in physics. Students exhibit many problematic aspects that are not strictly misconceptions, for example, misapplied procedures or lack of a piece of declarative knowledge. For this reason, Minstrell [7] refers to the diagnosed aspects as “facets of thinking,” influenced by diSessa’s Knowledge in Pieces [4]. For example, a common facet when describing forces is “bigger exerts more force” which interferes with a deep understanding of Newton’s Third Law.

Figure 1 shows an example DIAGNOSER question related to a set of facets on calculating average speed. In this case, the numerical response is used to diagnose one of a set of possible calculation errors (for example, often students forget to look at the initial time and position). Many questions are multiple-choice questions where the distractors correspond to commonly held facets. Consistency of student reasoning is often tested by asking the student to select, in a subsequent question, the statement that best corresponds to their reasoning. The full suite of DIAGNOSER questions and materials is available at www.diagnoser.com.

When used successfully in a classroom environment, the teacher will elicit students’ facets of thinking prior to instruction, and then, if those ideas do not conform to established scientific models and theories, the student will be given a chance to test those ideas with a series of experiments, or prescriptive activities. These will challenge the students’ beliefs and help them to move towards the target level of understanding. DIAGNOSER is used to identify and track those beliefs during instruction. In previous research [6, 7] we have seen that teachers who adopt DIAGNOSER and its associated teaching methods promote a deeper understanding of physics in their students. Thus, it is especially important to understand how a student is reasoning to identify the appropriate intervention. In the current system, the teacher obtains a report of facets diagnosed for each student on each question (approximately 6-10 questions in each set). Although the detailed information is crucial to help teachers understand the problems students may have, it is difficult for them to assimilate 5-10 facets diagnosed for each student in a class of 30 and determine how to appropriately challenge different groups of students.



Average Speed Set 1

John and his friends watch their radio-controlled car move along a straight path at their school track. John's friends mark the position of the car as it travels down the track. Some of the data are shown in the table below. Determine the car's average speed for the time interval shown in the table.

Position (m)	Time (sec)
10	2
15	3
18	4
21	5
23	6
24	7
25	8

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

□ meters/second

Figure 1. Screenshot of DIAGNOSER average speed question.

4. Entropy-Based Cluster Analysis

Cluster analysis is a commonly used technique to find and group items with similar attributes. The basic idea is to plot the items in a multidimensional space and group them into clusters according to the distance between them. When each attribute has a continuous value, it is easy to calculate the distance between attributes, and hence, between items. However, when attributes are categorical, the distance is less defined.

We are interested in how we might cluster student test responses to diagnostic questions such as those posed by DIAGNOSER. This allows us to identify groups of students with similar responses over a range of questions, to learn what facets of thought might be correlated and to simplify advice to teachers on appropriate instructional interventions. Let us consider students as items and their scores on a multiple-choice exam as a set of categorical-valued attributes. One approach to clustering categorical values is to consider the information content of each attribute. The mathematical definition of the information content of a signal state is the negative logarithm of the signal state probability:

$$I(P(s)) = -\log(P(s))$$

Intuitively, if the probability of getting a certain state is close to zero, the state transmits infinite information. If the probability of a certain state is 1, the information content is zero. However, we are usually concerned not with the entropy of a single state, but of a state within a distribution. We can calculate the Shannon entropy (H) for an entire column of values (in this case, answers to a single multiple choice question) by weighting each of the state entropies by the probability of seeing that state, and summing the results:

$$H(P) = -\sum P(s) * \log(P(s))$$

Thus, if all students select the same answer for a question, H=0. If answers to a four choice question are randomly distributed, H=2 (bits). For example, suppose the answer to question 1 is a multiple-choice answer that can assume the values A, B, C, and D. This question is basically an information channel that can assume four values, or two bits, of information. If all students randomly answer the question, the probability of each answer, or state, will be 0.25, and the minimum number of bits to convey the information in the column will be 2.

We can use the concept of entropy to place students into clusters according to their responses to a set of multiple-choice questions to minimize the entropy of the entire system. For example, suppose there is just one question and half of the students answer A, and half answer B. By our definitions above, we see that the entropy of this random variable is 1.0. If we can cluster the students into two groups, those who selected A and those who selected B, we can reduce the entropy of the system to zero.

More formally, we seek to minimize the *expected entropy*. We calculate the expected entropy by multiplying the entropy of each cluster (the sum of the entropies of each attribute) by the probability that an item falls into that cluster. Finding the optimal clustering is an NP-complete problem. Efficient implementations of entropy-based clustering have been described and implemented, such as COOLCAT [3] and LIMBO [2] and as we scale to large datasets we will need to evaluate such algorithms.

How do we evaluate the effectiveness of the clustering? The approach we have been using is to calculate the Categorical Utility function [5] which gives an idea of how predictive the clustering is, penalized by the number of clusters.

The Categorical Utility function is itself built up from a lower-order statistic, the Δ index introduced by Goodman and Kruskal. The Δ index is simply the decrease in the number of erroneous predictions of a variable's value that is obtained by knowing which cluster the case being predicted is in. This is best seen by an example. Suppose that on a particular two-choice question half the respondents say "Yes" and half say "No." Given this information, one expects to make an error in prediction half the time. Suppose further, though, that on the basis of other variables respondents are clustered into two groups, A and B. In cluster A 80% of the respondents answer "Yes" and 20% answer "no." Therefore the error rate for this variable, given knowledge that the case is in cluster A, is .20. The Δ index for the cluster is .50 - .20 = .30.

The Categorical Utility function adds Δ values for each attribute, and each cluster, and then introduces a penalty factor that increases with the number of clusters.

5. Patterns in the WASL

The statewide Washington (State) Assessment of Student Learning (WASL) examination is a typical, state-of-the-art educational assessment instrument, similar to those in use in many other states. We considered student results from the tenth grade WASL science examination given in Spring of 2003. This exam covers many areas of science, with only a few questions on each: the multiple choice questions have passed through the "Darwinian" IRT filter and should not be expected to reveal differences in individual thought. Some questions have been released and those questions and aggregate statistics are available at the Office of the Superintendent of Public Instruction website [1].

Analysis

The WASL assigns a student to one of four levels based on their overall score (1-4). Levels 1-2 represent students who are not proficient, and Levels 3-4 represent students who are proficient. We examined a 200-student sample from each level, to determine what items define the clusters in an information-theoretic sense at each level. There is good evidence to believe that students will apply different reasoning strategies as they are more proficient. Although there are open-ended and multiple choice questions, we limited our analysis to the 29 4-choice (plus no answer) questions. Open ended questions were not considered because we only had available the score assigned, and therefore knew the quality that the examiner had assigned the answer, but did not know what the student had said.

If the 29 4-choice questions were answered randomly the expected entropy would be 58. At the lowest level (Level 1) the entropy was nearly that great. Clustering reduced the entropy very little; the mean entropy within clusters (weighted by cluster size) was 48. This is an indication that students at the lowest level of proficiency were close to guessing.

Higher levels, where there is less noise in the data, reveal certain questions that stand out to define different clusters. Table 1 shows the entropy and distribution for a 4-cluster solution for Level 3; where the within cluster entropy is significantly lower. The Category Utility function is .197. Figures 2 and 3 show scree plots for Level 3 and Level 4. A scree plot is a plot in which questions are ordered by their mean Δ value after considering the cluster to which the case is assigned. The mean Δ value (ordinate) is then plotted against the order of the questions. This is analogous to a technique often used in factor analysis to determine the mean number of factors. If clusters were generated by random data the intervals between successive Δ values would be approximately the same. Discontinuities in Δ values indicate that the question with the higher value is making a substantial contribution to cluster formation.

We can see that at Level 3 question A stands out. (Actual question numbers have been removed to maintain security of the test.) A similar effect, along with another question, was found in the Level 2 clustering. Although we cannot reproduce the questions represented, they both require graph interpretation and understanding of velocity and acceleration: areas that are particularly problematic for students in physics. As we can see in Table 2, the clusters are largely defined by what distractors (b or c) the students selected besides the most popular selections (a and d). Patterns emerge from the confusion, but most students still do not correctly answer the question.

Cluster	H	Fraction of students in group
1	25.01	.20
2	24.56	.220
3	18.46	.365
4	27.85	.215

Table 1. Identified clusters for Level 3

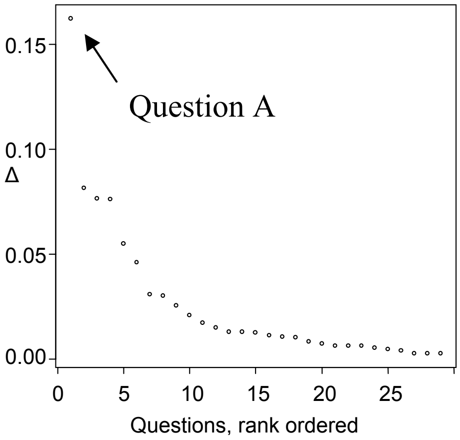


Figure 2. Scree plot for Level 3

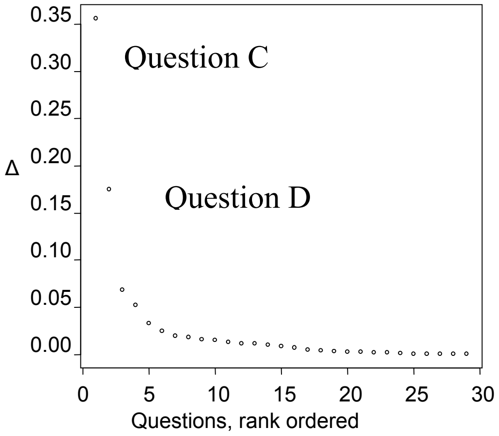


Figure 3. Scree plot for Level 4

	Selection				
Cluster	**a**	**b**	**c**	**d**	**blank**
1	26	2	4	12	0
2	25	0	0	23	1
3	47	0	4	13	0
4	19	8	0	16	0

Table 2. Count of responses to question 25 by cluster, Level 3

At Level 4 the optimal solution is a 6-cluster solution, and the entropy of the clusters is greatly reduced (ranging from 6 to 14). At this highest level, students are simply getting more questions right. However, the scree plot (Figure 2) shows that different questions emerge to drive the clustering. Question C is a question about geophysics. At level 2, students did not apply any consistent reasoning to this question, but at level 4 there was a coherent pattern to responses.

Discussion

The WASL is not designed to identify differences in student reasoning; however, at different levels we can identify that certain questions distinguish students more than others, and that those questions are different at various levels. Interestingly, some of the questions identified as problematic at levels 2 and 3 are treated at length in DIAGNOSER, and have large sets of facets associated with them to describe different student conceptions. This offers some small evidence for patterns of thought, even in such large scale assessments.

6. Patterns in a Diagnostic Assessment

The clustering approach we describe was designed to group students on assessments created to reveal patterns of thought. For this we used assessment data from an experiment carried out in a District A, in WA state, in June of 2004. Eight middle school science teachers received professional development on force and motion that consisted of five two-hour study group sessions between 11/03 and 2/04. Teachers attended between one and five of these sessions. Of these teachers, two used DIAGNOSER in subsequent instruction and six did not. It should be noted that the district in question had not, at that time, adopted a standard 7th grade force and motion curriculum and teachers had complete discretion over their instructional content.

We administered a 9-item assessment on Force and Motion topics to their students. These questions were written to correspond to item criteria for the state WASL. Topics covered included calculating speed and acceleration from graphs and tables and understanding the interactions of forces (these topics are also covered in DIAGNOSER). Each multiple choice answer for each question corresponded roughly to a different facet of student understanding in the DIAGNOSER system. To obtain more detailed information about student understanding, we asked the students to explain each answer in an open-ended question. We coded each of those explanations, in conjunction with their selected multiple choice answer, with a facet diagnosis, a certainty, and a consistency flag (i.e., were the students answering the question consistently with their stated reasoning). A total of 723 students took the test, 186 students in classes where teachers had used DIAGNOSER (DIAGNOSER users), and 537 students of teachers who had not (non-DIAGNOSER users).

Analysis

A simple analysis of scores (number questions correct) showed that students of DIAGNOSER users did approximately 15% better (effect size .82) than students of non-DIAGNOSER users. The difference persisted even when we normalized performance using the 7th grade math WASL scores (students do not take the science WASL, covering these topics, until 8th grade).

Thus, we know students exposed to DIAGNOSER are performing at a higher level. To answer the question of how, exactly, these students might be reasoning differently, we performed entropy-based clustering on groups of related questions from students of DIAGNOSER users and students of non-DIAGNOSER users. The clustering algorithm allows us to choose the number of clusters; we selected four as a start. Using the Categorical Utility function to assess the quality of different clusterings, we found the highest value to occur between 2-4 clusters (approximately .2).

We examined the distribution of answer selections for each cluster and identified the "modal" response. This is not limited to the single most frequent selection in the cluster, but the collection of multiple-choices responses that accounted for more than 75% of the students in that cluster, on a single question. For example, if answers in a cluster were evenly split for Q1 between a and b, the modal response for that cluster for Q1 would be a/b). Then we identified the reasoning diagnosed by hand for each question for all "modal responses". This reasoning characterizes the cluster.

The first four questions[1] asked students to interpret position vs. time and speed vs. time graphs and tables. The cluster breakdown for each group of students is shown in Table 4. We have ordered the clusters very loosely from top to bottom from "optimal understanding" to most problematic understanding (requiring the most instruction, theoretically, to bring to optimal). One interesting thing to observe from this data is that a cluster exists in the DIAGNOSER-user grouping that has an answer pattern corresponding to exactly what would happen if a student was quickly going through the exam without paying close attention to the questions. Questions 1 and 2 feature a position vs. time and a speed vs. time graph, respectively. If a student is not careful, he/she would answer Question 2 with answer B. If students truly do not understand differences in representation, they would probably make similar mistakes elsewhere in the test, and their flawed reasoning would appear elsewhere. In this case, it appears that this group understood how to read graphs but was not paying attention.

[1] We have made these questions available at www.facetinnovations.com/AAAI05.html.

Description	Modal Response	Diagnoser	Non-Diagnoser
Optimal understanding	A,D,B,A	37%	32%
Correct except that they read the graph in Q2 as a position v. time graph like Q1. Mistakes confusing position and speed did not appear in other questions.	A,B,B,A	19%	
Misread Q2 (slower means slowing down), misapplied reasoning to Q1.	C,A,B,A	22%	
Confusions of position and speed in graph and possibly table forms.	A,B/C,B/A,A		23%
Misread Q2 (slower), misapplied reasoning to Q1, and markedly confuse position and speed graphs or report position or distance instead of a speed.	C,A/B,A/B,A		23%
Truly have confusions regarding speed v. time and position v. time representations in multiple questions	C/D,D,B/A,A	21%	
Display a combination of literal interpretations of graphs (e.g. up on the graph is uphill) and confusions with position and speed.	C, D,A/B,A		21%

Table 4. Student reasoning on force and motion questions.

Description	Modal Response	Diagnoser	Non-Diagnoser
Optimal understanding, with some belief that forces may be imbalanced during the kick because it moves.	B,A,D/B	12%	
Understand that before the ball is kicked, gravity and the ground are acting on it. Believe in a "force of motion" propelling the ball after the kick. Forces equal during the kick.	B,C,D	27%	
May or may not understand that before the ball is kicked, gravity and the ground are acting on it. Believe in a "force of motion" acting on the ball after the kick. Forces equal during the kick.	B/A, C/B, D		21%
Understand that before the ball is kicked, gravity and the ground are acting on it. Believe in a "force of motion" propelling the ball after the kick. Believe forces imbalanced during kick because ball moves (non-diagnoser group also reasons that the bigger/heavier object exerts the greater force.)	B,C,B	32%	35%
Understand that before ball is kicked, gravity and the ground are acting on it. Believe in a "force of motion" propelling the ball after the kick. Believe forces imbalanced during kick because ball moves.	B,B,B/D	28%	
Most omit gravity as a force on the resting ball. Believe in "force of motion" propelling the ball after the kick, or unknown reasoning. Believe forces are imbalanced during the kick.	B/A,B,B		26%
Believe that before the ball is kicked, only gravity is exerting a force (inanimate objects do not exert forces). A "force of motion" propels the ball after the kick. Forces are imbalanced during the kick; bigger/heavier object exerts more force, or unknown reasoning.	A,C,B		17%

Table 3. Student reasoning on goalie kick questions.

Questions 5, 6 and 7 ask the student to describe the forces acting upon a soccer ball on the ground before it is kicked, during a kick, and after a kick. Results from clustering are shown in Table 5. In this case, no one cluster is singled out as the one with the optimal reasoning (that answer pattern would be B,A,D). Students who have that response pattern are grouped with those who also have some problematic beliefs surrounding force interaction pairs. Question 6, which asks about what forces act on the ball

Description	Modal Response	Diagnoser	Non-Diagnoser
Optimal understanding	C,B	36%	27%
Unknown reasoning, or correct answer for first question. Mostly unknown reasoning for second answer, with some believing in a sense of propulsion.	C/B, C	18%	
Reason in first question that the force acting on an object is proportional to the final speed and not the change in speed. In second question, split between correct reasoning, belief in a net force in the direction of motion, and unknown reasoning. *I believe this is all confusion related to friction…*	D/B, A/B (diag) D,A/B (non-diag)	27%	20%
Correct reasoning for the first question. In second, display belief that for an object to move at a constant speed, there must be an excess force in direction of motion.	C,A (diag) C/B, A (non-diag)	17%	28%
In the first question, split between correct reasoning and that the force acting on an object is proportional to the final speed and not the change in speed. For second question, mostly unknown reasoning and some belief in that if an object is moving at a constant speed, there must be a net force in direction of motion.	C/D, C		24%

Table 5. Student reasoning on forces and wagon

after it is kicked, has two choices (B: Gravity and the kick by the goalie, and C: Gravity and the force of motion) that are very similar in the sense that they are selected due to similar reasoning. In fact, selection of either usually results in diagnosis of the same underlying "force of motion" facet. We have no way in our clustering algorithm to indicate that these two responses are closer to each other than to either of the other responses to this question. Thus, some of the groupings are very similar and are driven inappropriately by responses B or C to question 6.

We can see from the first five rows that in general, the DIAGNOSER users understood the idea that before the ball is kicked, both gravity and the ground are acting on it. This facet does not characterize the beliefs of at least 17% of the non-DIAGNOSER users.

Questions 8 and 9 ask about what will happen to a wagon that is acted on by two forces. Results of clustering are shown in Table 6. The correct answers are C and B. We can see that 36% of DIAGNOSER users, vs 27% of non-DIAGNOSER users, were categorized by this optimal understanding. One difficulty in obtaining consistent clusters with this question is that many students assumed that the wagon was affected by a force of friction in addition to the force vectors indicated (the question did not specify "net" forces). Direct references to friction were coded as "unknown", and various ideas about friction might have corresponded to a variety of answers.

Discussion

Clustering was fairly effective in identifying patterns of reasoning among the student responses to the diagnostic assessment. However, the areas where a cognitive interpretation to clusters was not so straightforward raises some important issues. First, by its nature, entropy-based clustering on multiple-choice questions assumes that the distractors are cognitively equidistant from each other. They are not. We observed in the context of forces and interactions that there was not much difference between the belief in a stated "force of motion" and the belief that a kick can propel a ball after the initial contact. Second, we know before we begin that two of the clusters we are looking for are "optimal understanding" and "needs remedial attention". The optimal pattern does not necessarily emerge as a cluster when not enough students exhibit that reasoning. Alternatively, students who are missing prerequisite skills to answer the questions *and* those who are beginning to form reasoning strategies but are still inconsistent will both exhibit similar multiple choice answer patterns that have a high degree of noise. In other words, a student who cannot read graphs will answer randomly, and should be grouped with all other students who answer randomly, leading to a cluster with high entropy. We need a way to either model that noise or filter it.

Nevertheless, clustering can greatly simplify a diagnostic report for a teacher by placing students into a few rough categories. As formative assessment is used in the classroom to give the teacher feedback to help students move forward, this loose categorization is especially useful.

7. Conclusions and Future Work

Distinct patterns of thought along the ability scale were evident from analysis of both the WASL data and a diagnostic assessment designed to reveal student reasoning on topics of force and motion. Mining these patterns is a crucial step towards diagnostic assessment and educational intervention. However, although multiple-choice distractors can carry a lot of information about students' reasoning, there are limitations to what we can mine from them alone.

We are looking at model-based clustering using variable selection to better identify the questions that determine clusterings, and to obtain probabilities of students fitting the cluster. In addition, we are considering transforming the multiple-choice data based on consistency of student responses to more accurately calculate the "distance" between multiple choice selections.

Acknowledgements

I am extremely grateful to Jim Minstrell, Pamela Kraus, and James Minstrell for their help and encouragement on this project. This work was supported in part by NSF Applied Research Implementation grant 0138668.

References

[1] *Report Card*, Washington State Office of Superintendent of Public Instruction, last accessed April, 2005, available at http://reportcard.ospi.k12.wa.us.

[2] P. Andritsos, P. Tsaparas, R. J. Miller and K. C. Sevcik, *LIMBO: Scalable Clustering of Categorical Data*, *9th International Conference on Extending DataBase Technology (EDBT)*, 2004.

[3] D. Barbara, Y. Li and J. Couto, *COOLCAT: an entropy-based algorithm for categorical clustering*, *Proceedings of the eleventh international conference on Information and knowledge management*, ACM Press, McLean, Virginia, USA, 2002.

[4] A. A. diSessa, *Knowledge in pieces*, University of California, Berkeley, 1985.

[5] M. A. Gluck and J. E. Corter, *Information, uncertainty, and the utility of categories*, Artificial Intelligence, 40 (1985), pp. 11-62.

[6] E. Hunt and J. Minstrell, *A collaborative classroom for teaching conceptual physics*, in K. McGilly, ed., *Classroom lessons: Integrating cognitive theory and classroom practice*, MIT Press, Cambridge, 1994.

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

[8] R. J. Mislevy, R. G. Almond, D. Yan and L. S. Steinberg, *Bayes nets in educational assessment: Where do the numbers come from?* in K. B. L. H. Prade, ed., *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., San Francisco, 1999.

[9] C. Tatsuoka, *Data analytic methods for latent partially ordered classification models*, J Royal Statistical Soc C, 51 (2002), pp. 337-350.