Mining Diagnostic Assessment Data for Concept Similarity

TARA MADHYASTHA
University of Washington
madhyt@u.washington.edu
and
EARL HUNT
University of Washington
ehunt@u.washington.edu

Department of Psychology
Box 351525
Seattle WA 98195-1525

This paper introduces a method for mining multiple-choice assessment data for similarity of the concepts represented by the multiple choice responses. The resulting similarity matrix can be used to visualize the distance between concepts in a lower-dimensional space. This gives an instructor a visualization of the relative difficulty of concepts among the students in the class. It may also be used to cluster concepts, to understand unknown responses in the context of previously identified concepts.

Keywords: student modeling, diagnostic assessment, misconceptions, concept similarity, visualization, individual differences

---

1. INTRODUCTION

It is a popular idea in education that students come into a classroom with preconceptions about the material they are taught that can alter or interfere with their understanding of a topic. For example, when learning to read graphs of motion for the first time, students often interpret the graph literally, as they would a picture or a map. As students learn, they (hopefully) move from incorrect to correct concepts. It is recommended that teachers should check the extent to which students hold erroneous concepts throughout instruction, ideally to deliver personalized feedback to students. One such approach, developed by Minstrell, is called "Diagnostic instruction" [Hunt and Minstrell 1996; Minstrell 2001]. It is based on the idea of delivering multiple choice questions to students where each question (and corresponding responses) attempts to "diagnose" a particular type of thinking. In Minstrell's framework, these types of thinking are called "facets" and are catalogued by topic. Other researchers call the units of diagnosis misconceptions [Hamza and Wickman 2008], mental models [McCloskey 1993], or concepts [Hestenes 1992]. Some researchers posit that students switch between states that produce different types of misconceptions [Bao and Redish 2001; Huang 2003]. The multiplicity of theoretical frameworks underlies the general lack of agreement about the structure and role of incorrect concepts in the learning process.

The process of building taxonomies of common preconceptions is a laborious qualitative procedure. It is an iterative cycle of developing test questions, administering them to hundreds of students, and classifying open-ended responses into categories that appear most frequently. The most frequent responses then become multiple choice responses in revised items. However, there is no way to determine the relationships between these common misunderstandings. Are some responses very similar to each other? Do some represent higher or lower levels of understanding? What is the dimensionality of this space? These are important questions to answer in building a theory of how incorrect and partially correct ideas affect learning.

In this paper, we introduce a method for mining multiple-choice assessment data to determine the similarity of concepts represented by the multiple-choice responses. This method can be used to answer questions about the similarity of concepts and the difficulty of convincing students to change an erroneous concept. The remainder of this paper is organized as follows. In Section 2 we summarize the current state of mathematical models for diagnostic instruction, and motivate the need for this approach. In Section 3, we describe our methodology with a small example. Section 4 illustrates how our approach can be used to group concepts in order to gain insight into the underlying mechanisms that might cause certain types of incorrect thought. Section 5 shows how to use it to identify unlabeled misconceptions. Section 6 discusses future work.

2. BACKGROUND

Modern educational and psychological assessment is 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 (items). 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 weights of the 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. Assessments are made in an attempt to determine students' f_{ik} values, i.e. a student's place on the underlying ability scales.

IRT operates at the item level within a test. Consider the i^{th} item on a test. This item is assumed to have a characteristic difficulty level, B_i. Each examinee is assumed to have

skill level θ on the same scale. In the basic three parameter IRT model the probability that a person with ability θ will get item i correct is

$$P(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-B_i)}}{1+e^{Da_i(\theta-B_i)}}$$

where D is a constant scaling factor, a_i is an item discrimination parameter and c_i 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.

Neither of these commonly used models allow for idiosyncratic patterns of thought, where different people attack problems in different ways. More specialized models can describe mixtures of strategies [Huang 2003]. However, many educational theories are not easily fit to the assumptions of factor analytic or IRT models. Much of the motivation behind diagnostic assessment is to identify the different strategies that might change the relative order of difficulty of items.

Coming from a different modeling tradition, the intelligent tutoring community has developed techniques to automatically construct a representation of the knowledge domain, for the purposes of providing feedback or hints as a student works through a problem space. Falmagne et al. [1990] described a theory of knowledge spaces, with the goal of discovering the probabilistic relationship between items such that it is possible to identify the subgroup of problems that a student is likely to be able to solve. The reason for this departure from longstanding psychometric tradition was that it seemed more grounded in the immediate needs of a tutoring system. A simplification of knowledge spaces is partial order knowledge structures [Desmarais et al. 1995]. In both cases, probabilistic relationships can be estimated to form a network structure. A strength of this approach is the flexibility of the underlying knowledge space. However, knowledge structures do not provide estimates of the distances between knowledge states and do not incorporate the states associated with failed attempts at problems (although they could be included in practice as correct responses to problems designed specifically to elicit these ideas). These incorrect conceptions are considered to be important in many domains, where progress requires specifically addressing the misconception.

Several efforts have taken a non-cognitive view of response spaces in order to provide feedback. Fosatti and colleagues constraint-based modeling to compare a student response against a set of constraints, where violated constraints correspond to gaps in knowledge or incorrect understanding [Fosatti 2008]. This approach precludes the need to construct a complex model of knowledge. Stamper and Barnes exploit the idea of consistency of student actions in the aggregate (e.g., that if many students do something, there is likely to be some good reason behind it) in a tutor based on a Markov decision process [Stamper and Barnes 2009]. The most frequent and useful steps in problem solving that other

students have taken are used to provide hints for other students.

The problem of how best to mathematically model a knowledge space is open, and the answer may be domain-dependent. There is evidence suggesting that in fact, facets (fine-grained correct, partially correct, and incorrect understandings) may have a structure to them in some domains that can be modeled using a partial credit model [Wright and Masters 1982]. Using this model, multiple choice responses are ordered in difficulty on a linear scale, allowing one to rank students by ability based on their responses. This implies that the relative difficulty of items in some interesting domains may indeed be the same for all students [Scalise et al. in press]. Thus, modeling can be improved by identifying this linear structure of concepts. Each item response would have its own difficulty on a linear scale, providing a clear measure of student and classroom progress, e.g., a learning progression where content is mapped to an underlying continuum [Wilson 2004; Wilson and Sloane 2000]. But building this knowledge representation is an extremely large endeavor, especially in subject areas where little research has been done into the ideas students have before instruction that affect their understanding, or what dimensional structure is appropriate to represent them. There is the danger that IRT scaling would result in inadvertedly discarding items that do not fit the model but might reveal additional interesting information about student thought. In contrast, automatically constructed knowledge spaces may lead to overestimation of knowledge states.

We feel that there is a call for an exploratory data-mining approach that can be used in qualitative research leading up to development of concept taxonomies, and implemented in tools that can be used in the classroom for understanding the results of diagnostic assessment. Although we demonstrate this method on diagnostic questions related to physics instruction, following Minstrell's approach [Hunt and Minstrell 1996; Minstrell 2001] the technique can be used to identify consistent, but possibly erroneous, reasoning in any data bank containing responses to multiple choice questions. Indeed, since the technique identifies clusters of patterns in multiple choice responses, regardless of content, it could be applied outside of education, for instance to surveys of political opinions or questionnaires about health practices.

3. METHOD

Our intuition is that if we give students a pre-test and a post-test with nearly identical items, especially with little instruction in between, the changes in aggregate response can be used to construct a measure of similarity between concepts. To operationalize this intuition, we identify a metric of agreement, drawing from the literature on the measurement of inter-rater agreement. Typically such metrics are used to quantify agreement among two independent raters who assign each subject into a category. For

example, two raters might observe a set of people and describe each of them as “tense”, “neutral”, or “happy”. In that case, we would use the concept of inter-rater agreement to quantify how well the two raters agreed on the coding. Ideally both raters would agree on their descriptions of each person, producing high inter-rater agreement. If, however, one rater always says a subject is “tense” when the other rater says the subject is “neutral”, there is low agreement, and the concepts of “tense” and “neutral” are easily confused, and therefore similar to one another. In our case, two “ratings” are made by students who assign an item (question) to a category (response concept) at two time points. Therefore student consistency is an analog of inter-rater agreement.

Consider a pair of questions as shown in Figure 1. This question pair is designed to elicit ideas that 7th graders use to interpret a speed versus time table. We note that the items differ in content characteristics (e.g. a runner versus a car, and their corresponding speeds) but they are conceptually the same question. We say that this question pair is *isomorphic.* If a person is using consistent abstract reasoning his or her response to one question in an isomorphic pair should predict the response to the second question.

The student should note from the table in Figure 1 that the speed is decreasing at the specified time. Therefore, the correct answer is C*,* slowing down. Students often answer A, that it is speeding up, when looking at the general trend of speeds in the table. They justify answer B (constant speed) by saying that the speeds are increasing at a constant rate, or that for that one second, the speed is constant. Answers for D (not moving) generally indicate that the student believes an object can’t be moving at any specific instant.

These questions appeared on a pre-test and post-test administered to 7th graders in a medium sized suburban school district in 2005. 1026 students took the pre-test and 981 students took the post-assessment, and 925 students took both the pre-assessment and post-assessment. Some assessments were given on computer and some on paper, so not every student answered every question on the paper version. Between assessments, students had between 1 and 12 weeks of instruction on topics in motion and forces using the Full Option Science System (FOSS) Force and Motion middle school curriculum. In this situation we are interested in the extent to which choice of one response on a pretest predicts another on the post test. In an educational application this provides an evaluation of the effect of instruction, conditional upon the prior ideas of the student. Using our earlier definition, though, the educational question is isomorphic to, for instance, an evaluation of the effect of an advertising campaign. Note that our educational question goes beyond conventional “right or wrong” scoring, where the only interest is in comparing the probability of a correct answer on the pretest and the posttest.

Pre-test

In the table at right, the speed versus time data for a runner is given for a 6 second trip. Use this data to answer the following three questions.

4) Which description best fits the motion of the runner at t = 1 second?

A. Speeding up
B. Constant speed
C. Slowing down
D. Not moving

Briefly explain your reasoning in the space provided.

Speed vs. Time for a Runner

Speed (m/s)	Time (sec)
3	0
2	1
1	2
1	3
1	4
2	5
3	6

Post-test

In the table shown, the speed versus time data for a car is given for a 6 second trip. Use this data to answer the following two questions.

8) Which description best fits the motion of the car at t = 1 second?

A. Speeding up
B. Constant speed
C. Slowing down
D. Not moving

Briefly explain your reasoning in the space provided.

Speed vs. Time for a Car

Speed (m/s)	Time (sec)
80	0
78	1
76	2
76	3
76	4
80	5
84	6

Figure 1. Questions from 7th grade pre and post test.

Table 1 shows the proportions of students who answered with each combination of responses on the pretest and the posttest. To consider the agreement among two concepts, e.g. the response C on the pretest and on the posttest, we collapse the categories into "Selected C" and "Selected anything else" to form the table of proportions given in Table 2.

Table 1. Responses to questions 8 and 4 (920 students).

	Posttest				
Pretest	A	B	C (correct)	D	Total
A- speeding up	0.01	0.01	0.06	0.00	0.09
B - constant speed	0.01	0.02	0.06	0.00	0.09
C- slowing down (correct)	0.02	0.08	0.69	0.01	0.80
D- not moving	0.00	0.00	0.01	0.00	0.01
	0.00	0.12	0.83	0.01	1.00

Table 2. Measuring agreement on a single concept.

	General				C- slowing down		
	Posttest				Posttest		
Pretest	Given Concept	All Others	Total	Pretest	C	All Others	Total
Given concept	*a*	*b*	p_1	c	0.69	0.11	0.8
All Others	*c*	*d*	q_1	All Others	0.14	0.06	0.2
Total	p_2	q_2	1		0.83	0.17	1

From a table such as Table 2, many measures of agreement have been proposed. We use the proportion of specific agreement, p_s [Fleiss et al. 2003] :

$$p_s = \frac{2a}{2a+b+c}$$

In the example above, the calculation of agreement on C is:

$$p_s = \frac{2\times.69}{2\times.69+.11+.14} = .85.$$

The proportion of specific agreement for a concept X can be interpreted as the probability of selecting X on the pre-test and the post-test, divided by the probability of choosing it on either. Similarly, there are two values that can be calculated for the proportion of specific agreement for different concepts X and Y. The first is the probability of choosing both X on the pre-test and Y on the post-test, divided by the probability of choosing X on the pre-test or Y on the post-test. The second is the probability of choosing both Y on the pre-test and X on the post-test, divided by the probability of choosing Y on the pre-test or X on the post-test

We want to compare p_s to the probability of agreement by chance. This is done by calculating kappa ($\hat{\kappa}$), or the ratio of the observed excess beyond chance and the maximum possible excess. This metric has been used in such areas as ecommerce [Ichise et al. 2003] and life sciences [Kirsten et al. 2007].

$$\hat{\kappa} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

The main difference between p_s and $\hat{\kappa}$ is that the value of *d*, or agreeing on something besides the concept of interest, is not considered. This makes sense for this application, when selection of a particular response might be relatively rare, and failure to select it would increase the measure of agreement. Note that $\hat{\kappa}$ will be biased by a large value of *d*, and so may not be particularly informative (especially if the probability of expected chance agreement is low, as is the case in many diagnostic assessment questions). For this reason, in all examples in this paper, we have found p_s to be a more useful statistic of agreement than $\hat{\kappa}$.

Using the method above, we can calculate for questions 4 and 8 the values of p_s for all concepts, as shown in Table 3.

Table 3. Proportion of specific agreement for questions 4 and 8.

	Posttest			
Pretest	A-speeding up	B-constant speed	C-slowing down (correct)	D-not moving
A - speeding up	.19	.13	.14	.02
B - constant speed	.10	.20	.14	.04
C - slowing down (correct)	.05	.17	.85	.02
D - not moving	0	.07	.02	.10

If all students selected the same response both times, p_s =1. A value close to zero indicates that few students selected the same response twice. Table 3 shows that the response C was selected most consistently, which is reassuring because it is the correct response. A persistent misconception, if diagnosed by this question, would have a high value of p_s.

It is instructive to examine the similarities between other responses and the correct response. The proportion of similarity between C and B is .17, which is just slightly more than the proportion of similarity between B and C (.14), indicating that there is confusion between these two concepts that was not resolved adequately with instruction. Indeed, the increase in correct responses was only on the border of statistical significance (t(920) = .054). Asymmetries in the matrix indicate areas where the similarity between concepts was either increased or reduced during instruction. For example, the similarity between C (pretest) and A (posttest) is .05, which is smaller than the similarity between a (pretest) and C (posttest), which was .14. Instruction had the effect of moving students from

response A to the correct response.

It is helpful to visualize the similarity between concepts. When item responses can be fit to an appropriate IRT model, each response can be mapped to a certain difficulty on a linear scale. A commonly used technique for viewing this is a Wright Map. This is a graphic where item responses are ordered from easiest to hardest (by the B_i parameter), and a histogram of respondents is aligned to the difficulty of items. This allows an instructor to determine how many students have a 50% chance of answering each item correctly.

Analogously to ordering items by difficulty, we can use the technique of multidimensional scaling (MDS) to visualize the distances between concepts in a two-dimensional space. MDS is a statistical procedure that rearranges points in a multidimensional space (e.g., an arbitrary matrix of distances) into a lower dimensional space, preserving the distances between points as much as possible so as to reproduce the original matrix. Goodness of fit is measured by a stress measure, e.g., the sum squared deviations of the reproduced distances from the original distances.

First, we convert a similarity matrix as given in Table 3 to a distance matrix by averaging the matrix with its transpose, setting the diagonal values to 1 (ignoring information about the similarity of a concept to itself) and subtracting the similarity of the other cells from 1. This means that the similarity of two different concepts X and Y is the average of their proportions of specific agreement whether X is on the pre-test and Y on the post-test, or vice versa. The average is particularly useful when there has been little conceptual change between testing occasions, as evidenced by fairly symmetric values for specific agreement across different concepts on the pre-test and post-test.

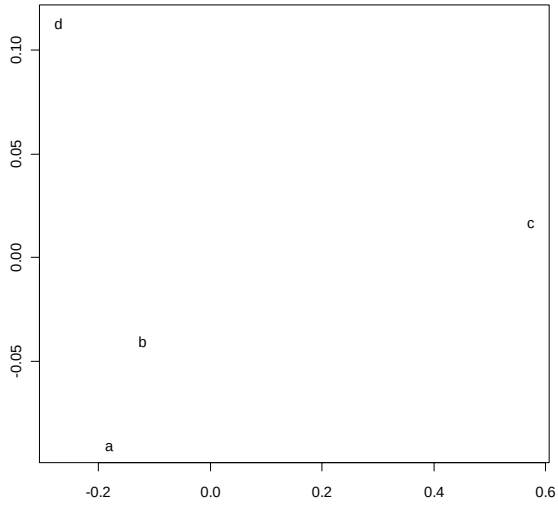


Figure 2. Multi-dimensional scaling representation of distance matrix for questions 4 and 8.

Figure 2 shows the multidimensional scaling solution obtained for the average of the upper triangular and lower triangular matrices of Table 3 (using the `cmdscale` function in R [R DEVELOPMENT CORE TEAM, 2008]). Our choice of two dimensions is arbitrary, for visualization purposes, but the goodness of fit statistic is .986 (highly satisfactory), indicating that two dimensions is sufficient to reproduce the distances. This figure shows that the correct answer C is quite far from all the incorrect responses. Responses A and B form a small cluster, and response D occurs by itself. This suggests that in this population of students, the concepts represented by responses A and B are similar, and D represents a different kind of misunderstanding.

4. EXAMPLE 1: GROUPING CONCEPTS

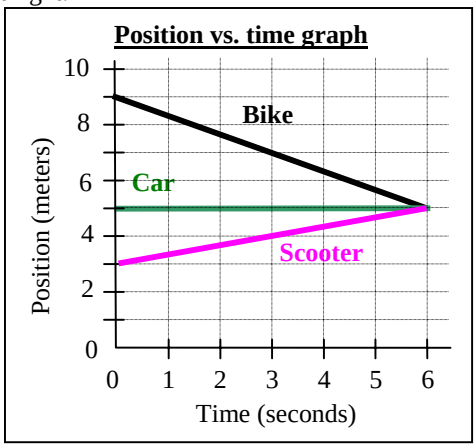
One of the controversies in educational research on the diagnosis of concepts is whether students form self-consistent incorrect mental models, or whether their knowledge is fragmented, or some combination of the two [Elby 2000]. We show how our method might be used to gain insight into the underlying reasoning strategies behind selections of responses.

Figure 3 shows two paired questions from the same pre and post assessments given to 7^{th} graders in 2005. In this case, students were asked to choose the multiple choice response to explain their reasoning. Table 4 shows the reasoning coding (using Minstrell's facet code) for this post-test question pair. Many possible combinations of responses, not shown in the table, are logically inconsistent, defying assignment of a reasoning code. In this example, we show how similarity of responses can reveal underlying reasoning difficulties for those responses that are logically consistent.

First, we recode the data according to the diagnosed reasoning pattern, to account for the fact that the questions are slightly different. All unknown pairs receive the code "Unk". Note that wherever a student selects "g" for the second question, their reasoning might or might not be consistent with their response to the first question, but it is still coded as "Unk". On the posttest, 76.3% of students selected a response pattern that was consistent with a particular reasoning pattern.

Pre-test

The position versus time for three objects, a bike, a car, and a scooter, have been graphed at right.

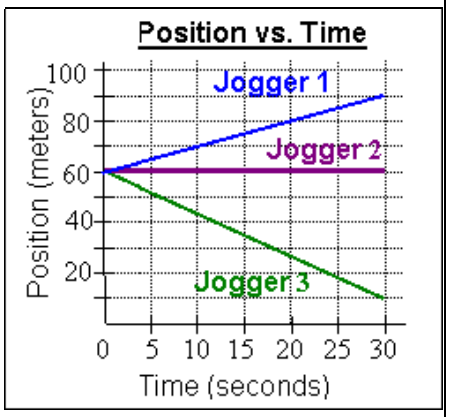


1) Which object is going the fastest (on average) during the six seconds of motion?
A. The Bike
B. The Car
C. The Scooter
D. They all traveled at the same speed.

2) Which statement below best matches the reasoning you used in answering the previous question?
A. The scooter's numbers are increasing.
B. The bike went the furthest in 6 seconds.
C. All the bike's numbers are bigger than the car or the scooter.
D. The bike is going downhill, so it is always getting faster and faster.
E. They are all at 5 meters at time 6 seconds.
F. The line for the car is the shortest
G. Other. (Please describe in the space provided.)

Post-test

The position versus time for three different joggers have been graphed at right.



3) Which jogger is going the fastest (on average) during the 30 seconds of their motion?
A. Jogger 1
B. Jogger 2
C. Jogger 3
D. They are all moving at the same speed

4) Which statement below best matches the reasoning you used in answering the previous question?
A. Jogger 1's numbers are increasing.
B. Jogger 1's numbers are bigger than 2 or 3's.
C. The line for Jogger 2 is the shortest
D. Jogger 3 went the furthest in 30 seconds.
E. Jogger 3 is going downhill, so she is always getting faster and faster.
F. They all finished at the same time.
G. Other. (Please describe in the space provided.)

Figure 3. Calculating speed from position vs time graph.

Figure 4 shows the multidimensional scaling solution obtained from applying the technique described above to this set of responses for the average of the upper triangular and lower triangular matrices of the distance matrix (N=920). Goodness of fit was .869, suggesting that two dimensions reproduced the distances well. Note that there are essentially three groups of ideas. The first, described by DS90, DS91, PD40, and PS70, may all be seen as variants of a primitive reasoning strategy that Elby calls "What you see is what you get" [Elby 2000]. In our context, this means looking at the graph and making

some kind of literal interpretation. Although only DS90 and DS91 refers specifically to the literal reading of the graph as a map of motion or a depiction of terrain, the question pair 3D-4F, corresponding to PD40 (student does not distinguish between position and speed), may be seen to draw on visual elements of the graph for the corresponding reasoning, noting that all the lines stop in the same place. Similarly, 3A-4B, corresponding to PS70, can also be viewed quite clearly as a manifestation of "higher is faster". This supports a theory that student reasoning is fragmented, as opposed to consistent and theoretical, because there is so much confusion between these seemingly different responses. An analysis of this kind might give insight into the underlying problems that students are having by revealing their similarity to each other.

It is interesting that DS80 is quite distinct from these other reasoning strategies. Students who respond with this question pair may be confusing the position versus time graph with a speed versus time graph. This mistake is very common, even among college students [McDermott et al. 1987]. We have found it to be extremely prevalent among 7th and 10th graders, even those who have mastered most of the material.

Table 4. Description of responses to calculating speed from position versus time graph.

Cluster title	Response Pattern	Facet Code	Description of facet of student thinking
Determining Speed	**1A-2B 3C-4D**	**DS02 (correct)**	Given position vs. time data, student correctly describes and determines the speed of an object moving uniformly.
	1C-2A 3A-4A	DS80	Student confuses position vs. time and speed vs. time graphs or data tables.
	1B-2C 3B-4F	DS90	Student views a position or speed graph as a map of the actual motion.
	1A-2D 3C-4E	DS91	Student interprets an upward (or downward) sloping graph to mean the object is going up hill (or downhill).
Position and Distance (PD)	1D-2E 3D-4F	PD40	The student does not distinguish between the ideas of position and distance.
Position and Speed (PS)	1A-2C 3A-3B	PS70	The student does not distinguish position and/or distance from speed.

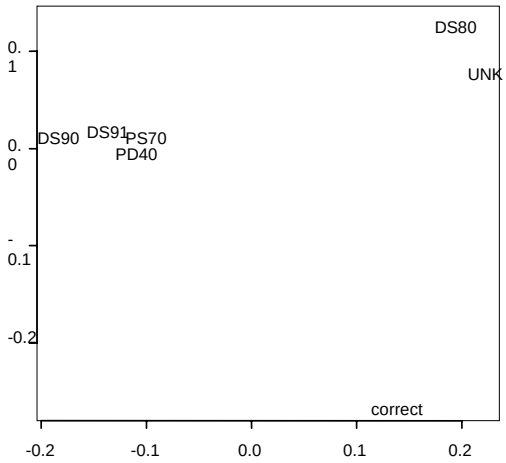


Figure 4. Multidimensional scaling representation of similarity of concepts for determining speed from position vs time question pair (920 students).

5. EXAMPLE 2: CLUSTERING AND IDENTIFYING CONCEPTS

We turn our attention to using our measure of concept similarity to identify concepts that have not previously been identified by pedagogical content experts.

This example uses data from the same 2005 force and motion assessment (N=906). In this example, we combine responses to three isomorphic "choose all that apply" type questions to identify common strategies for reasoning about forces and motion in the context of moving an object on a frictional surface. Figure 5 shows the first of the questions in this scenario. This question shows a hand pushing a block along a table, and asks the student to describe the relative forces of the hand and friction that are required to make the block speed up. The second and third questions of the scenario are identical, except that they ask about the relative forces required to make the block move at a constant speed or slow down, respectively.

The correct response pattern is *abc-d-e* (i.e, responses *a*, *b* and *c* checked on the first question, response *d* checked the second question, and response *e* checked on the third question). To get the object to speed up, force should be larger than friction but may be constant or changing. To get the object to move with constant speed, the force should be equal to friction. To make the object slow down, the push must be smaller than friction. To be fully correct, a student should notice that all options *a*, *b*, and *c* will result in the object speeding up. However, this subtlety is extremely difficult for 7th graders, and indeed no one answered with the correct response pattern on the posttest. When scoring this question, we gave students credit for any answer to the first question of the scenario that included *c* alone, or c and one or both of *a* and *b*. However, to distinguish these

responses, we label the response pattern *abc-d-e* **correct**, *bc-d-e* **correct1** and *ac-d-e* **correct2**.

The pattern selected with the highest frequency (12.9%) of the pre-assessment for the series of three questions was *a-d-e* (increasing/equal/smaller), which we will call **impetus**. Here, the student believes that to get the block to speed up, it needs an increasing force. Once it is in motion, the force need only be equal to friction to keep it going with constant speed. To make it slow down, the force should be smaller than friction. This pattern of responses is incomplete, as described above. Any of the selections that state the push is larger than friction (increasing and decreasing) will result in the object speeding up. One reason students may be drawn to the "larger and increasing" option is that they may equate speeding up with increasing acceleration, implying increasing force. This is consistent with the findings of [Trowbridge and McDermott 1981] where ideas of speed and acceleration are often interchanged.

The second most prevalent pattern (9.6%) was *a-c-b* (increasing/larger/decreasing), which we call **prop**. This type of response has been well documented in research on older students [McDermott and Redish 1999]. Students believe that to get the block to speed up, the push needs to be larger than friction and increasing. To make it move with constant speed, the force should be larger than friction. To make it slow down, the push should be larger than friction but decreasing. This pattern of responses is consistent with Clement's "motion implies force" idea [Clement 1982] and the "common sense belief system about motion" described by Champagne *et al* [1980] that includes the idea that "the magnitude of the velocity is proportional to the magnitude of the force: any acceleration is due to increasing forces." Basically, students are responding with the notion that the net force is proportional to the speed of the object [Viennot 1979].

Of similar frequency on the pre-assessment (9.6%) is the pattern *a-d-b* - that the force by the hand must be larger than friction and increasing to make the object speed up, equal to friction to make it move with constant speed, and larger than friction and decreasing to make it slow down. This pattern might be similar to the **prop** pattern, because in both patterns the force by the hand is related to changes in speed, but it was not previously identified. We call this response pattern **NF2**.

Two patterns occurred with modest frequency (2.9% and 3.5%) only after instruction: *ac-c-bc* (**X1**) and *ac-c-bc* (**X2**). These seem to be variants of the **prop** pattern, but they were not previously identified.

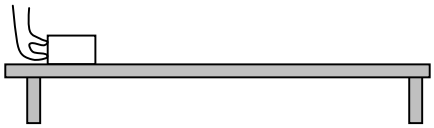
We recoded each response chosen by fewer than 10 students on both the pretest and the posttest to "**other**". This approach exploits the intuition that if a response pattern represents a coherent line of reasoning, several students should have the same response pattern. This principle is similar to the principle behind Consensus Based Measurement

[Legree et al. 2005] used to determine correct answers to situational questions from non-expert examinees. We feel confident that responses coded as **other** reflect a fairly low level of understanding.

We constructed a distance matrix using these identified response patterns in the pretest and the posttest (correct, correct1, correct2, impetus, prop, NF2, X1, X2 and other) and any additional patterns that students shifted from or to using the method described in Section 3.

Pre-test

1) Once the block is moving to the right, which of the following comparisons between the force by the hand and the force of friction will get the block to **speed up**? Select all that apply.



The force by the hand is…

a. …larger than friction and increasing.

b. …larger than friction and decreasing.

c. …larger than friction.

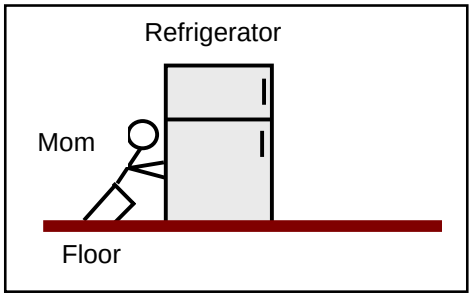
d. …equal to friction.

e. …smaller than friction.

f. Other. Explain in the space below.

Post-test

1) For 3 seconds the refrigerator **speeds up**. Which of the following comparisons between the force of the mom's push and the force of friction will get the refrigerator **to speed up**? Select all that apply.



The force by the mom is…

a. …larger than friction and increasing.

b. …larger than friction and decreasing.

c. …larger than friction.

d. …equal to friction.

e. …smaller than friction.

f. Other. Describe in the space provided

Figure 5. Question 1 of 3 question sequence on identifying forces.

Figure 6 shows the resulting distance matrix, clustered according to a hierarchical clustering algorithm with complete linkage (R `hclust`). We choose this approach to visualization of the concept distances instead of MDF because it is easier to view the relationships between the concepts. At each step in the clustering, patterns that are closest together are merged. Therefore, patterns that are linked higher up in the tree are more

distant from each other than those that are connected at lower levels in the tree. This allows us to confirm the relationship of response patterns that we have hypothesized should exist, and to interpret other frequent response patterns in the context of their relationships to the previously identified patterns.

The first thing to notice is that **impetus** reflects a fairly primitive understanding, by its proximity to **other**. The response e-d-b occurs with very low frequency on the pretest (11 responses) and only once on the posttest. Second, we note that **NF2**, despite its superficial similarity both in frequency and pattern to **prop**, is quite distinct from that pattern. **X1** is a much closer variant to **prop**, followed by **X2**. We believe that this illustrates a novel variant of incorrect understanding that may arise from a different kind of reasoning than force proportional to speed. This would be something that a pedagogical content expert might explore.

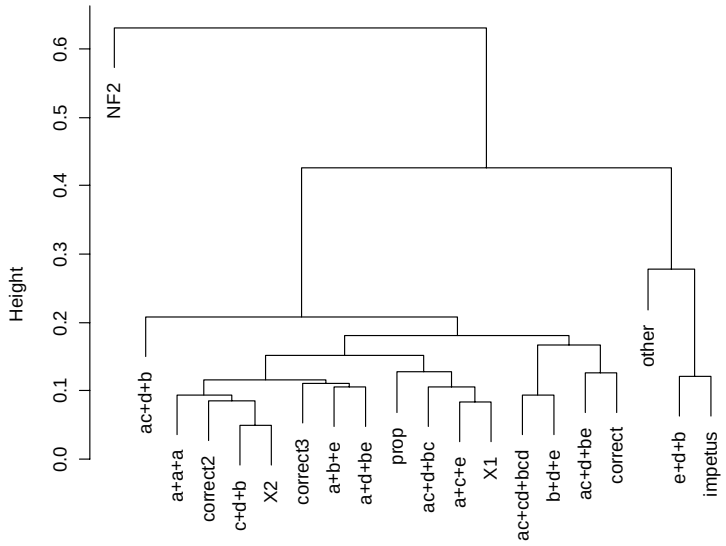


Figure 6. Tree produced by hierarchical clustering of distance matrix for most responses to forces to explain motion scenario.

Finally, we note that students who stumbled upon variants of the correct answer probably did not actually understand the underlying concepts; their responses are much closer to prop, X1 or X2 reasoning. This is an example of the population-specific nature of this approach to calculating distance; among a group of experts, one should expect a different kind of knowledge structure.

6. CONCLUSIONS AND FUTURE WORK

We have presented a technique for computing distances between concepts, as represented by selections on multiple-choice questions, within a specific population. Distances may be mapped, or used as input to a clustering algorithm. This technique is of interest as a method to understand how knowledge about a specific domain might be structured. Specifically, we have shown its utility as a way to examine the relationship between concepts that have been identified by the literature. We have also demonstrated how it can be used to understand what uncoded responses might mean, through their proximity to previously targeted concepts.

A limitation of this approach is that it relies upon student responses to isomorphic questions with little learning in between. In this context these are multiple choice or "choose all that apply" format questions, but any response that can be mapped to a categorical choice is appropriate for application of this method. However, the questions must have the same form, so that a student should be theoretically expected to answer the same way if presented with the questions in sequence. Practically, this means that the questions should be of the same difficulty and should measure achievement of the same skill in the same way. These kinds of questions could be generated from a parameterized computer model [Graf 2008].

However, unexpected conceptual "distance" could occur when supposedly similar questions and responses actually tap into different underlying constructs. For example, in the context of physics, surface features (e.g., the objects involved, directions of actions, the damage that occurs) often invoke different types of student thought. Among beginning physics students, a question about the forces involving two cars of equal size colliding will likely elicit extremely different descriptions of the interacting forces than a question about a car hitting a fly. Questions about these situations may be isomorphic to experts but not to beginners. As we have emphasized earlier, concept distances calculated using this method are population-dependent, which is another way of gauging movement from novice to expert understanding. The expected distance between the same concepts is different for experts and novices; therefore, one could compare distance matrices of novices to those of experts to gauge movement towards expertise.

The method accommodates visualization of the distances between concepts in a complex multidimensional space. However, related work [Scalise, Madhyastha, Minstrell and Wilson in press] indicates that related groups of concepts in physics may have a more linear structure. Evidence for this is that the distances found through this data mining technique can be satisfactorily reduced (e.g., by a multidimensional scaling solution) to a lower dimensional space, or clustered to a few groups.

We have exploited the concept of group consistency of response to uncover a general

concept structure within a group of students (e.g. a grade level or a classroom). Further research would examine to what degree individuals hold consistent but incorrect concepts. This would help us better understand how to target feedback to incorrect concepts within a specific problem domain.

ACKNOWLEDGEMENTS

The authors thank Jim Minstrell and Pamela Kraus at FACET Innovations for providing access to the data used in this paper.

REFERENCES

BAO, L. and REDISH, E.F. 2001. Concentration analysis: A quantitative assessment of student states. *American Journal of Physics 69*, S45-S53.

CHAMPAGNE, A., KLOPFER, L.E. and ANDERSON, J. 1980. Factors Influencing the Learning of Classical Mechanics. *American Journal of Physics 48*, 1074-1079.

CLEMENT, J. 1982. Students' Preconceptions in Introductory Mechanics. *American Journal of Physics 50*, 66-71.

DESMARAIS, M.C., MALUF, A. and LIU, J. 1995. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction 5*, 283-315.

ELBY, A. 2000. What students' learning of representations tells us about constructivism. *The Journal of Mathematical Behavior 19*, 481-502.

FALMAGNE, J.-C., KOPPEN, M., VILLANO, M., DOIGNON, J.-P. and JOHANNESEN, L. 1990. Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review 97*, 201-224.

FLEISS, J.L., LEVIN, B., PAIK, M.C. and FLEISS, J. 2003. *Statistical Methods for Rates & Proportions*. Wiley-Interscience.

FOSATTI, D. 2008. The role of positive feedback in intelligent tutoring systems. In *Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume)* Association for Computational Linguistics, Columbus.

GRAF, E.A. 2008. Approaches to the Design of Diagnostic Item Models Educational Testing Service, Princeton, NJ.

HAMZA, K.M. and WICKMAN, P.-O. 2008. Describing and analyzing learning in action: An empirical study of the importance of misconceptions in learning science. *Science Education 92*, 141-164.

HESTENES, D. 1992. Force Concept Inventory. *Physics Teacher 30*, 141-158.

HUANG, C.-W. 2003. Psychometric Analysis Based on Evidence-Centered Design and Cognitive Science of Learning to Explore Student's Problem-Solving in Physics, University of Maryland.

HUNT, E. and MINSTRELL, J. 1996. Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education 2*, 123-162.

ICHISE, R., TAKEDA, H. and HONIDEN, S. 2003. Integrating multiple internet directories by instance-based learning. *In: Proceedings of the eighteenth International Joint Conference on Artificial Intelligence. (2003*, 22--30.

KIRSTEN, T., THOR, A. and RAHM, E. 2007. Instance-Based Matching of Large Life Science Ontologies. In *Data Integration in the Life Sciences*, 172-187.

LEGREE, P., PSOTKA, J., TREMBLE, T. and BOURNE, D. 2005. Applying Consensus-Based Measurement to the Assessment of Emerging Domains, Army Research Institute for the Behavioral And Social Sciences.

MCCLOSKEY, M. 1993. Naive theories of motion. In *Mental Models*, D. GENTNER and A.L. STEVENS Eds. Lawrence Erlbaum, Hillsdale and London, 299-324.

MCDERMOTT, L.C. and REDISH, E.F. 1999. RL- PER1: Resource Letter on Physics

Education Research. *American Journal of Physics 67*, 755--767.
MCDERMOTT, L.C., ROSENQUIST, M.L. and ZEE, E.H.V. 1987. Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics 55*, 503-513.
MINSTRELL, J. 2001. Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In *Designing for Science: Implications for Professional, Instructional, and Everyday Science*, K. CROWLEY, C.D. SCHUNN and T. OKADA Eds., Mawah, NJ.
R. DEVELOPMENT CORE TEAM, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org, Vienna, Austria.
SCALISE, K., MADHYASTHA, T., MINSTRELL, J. and WILSON, M. in press. Improving Assessment Evidence in e-Learning Products: Some Solutions for Reliability. *International Journal of Learning Technology (IJLT)*.
STAMPER, J. and BARNES, T. 2009. An Unsupervised, Frequency-based Metric for Selecting Hints in an MDP-based Tutor. In *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings*, T. BARNES, M. DESMARAIS and S. VENURA Eds., Cordoba, Spain.
TROWBRIDGE, D. and MCDERMOTT, L.C. 1981. Investigation of Student Understanding of the Concept of Acceleration in One Dimension. *American Journal of Physics 49*, 242-253.
VIENNOT, L. 1979. Spontaneous Reasoning in Elementary Dynamics. *European Journal of Science Education 1*, 205-221.
WILSON, M. 2004. *Constructing Measures: An Item Response Modeling Approach*. Lawrence Erlbaum.
WILSON, M. and SLOANE, K. 2000. From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education 13*.
WRIGHT, B.D. and MASTERS, G.N. 1982. *Rating Scale Analysis*. Pluribus.