Concentration Analysis:
A Quantitative Assessment of Student States

Lei Bao
Department of Physics, Kansas State University, Manhattan, KS66506

Edward F. Redish
Department of Physics, University of Maryland, College Park, MD 20742

Multiple-choice tests such as the Force Concept Inventory (FCI) provide useful instruments to probe the distribution of student difficulties on a large scale. However, traditional analysis often relies solely on scores (number of students giving the correct answer). This ignores what can be significant and important information: the distribution of wrong answers given by the class. In this paper we introduce a new method, concentration analysis, to measure how students' responses on multiple-choice questions are distributed. This information can be used to study if the students have common incorrect models or if the question is effective in detecting student models. When combined with information obtained from qualitative research, the method allows us to identify cleanly what FCI results are telling us about student knowledge.

I. Introduction

Both physics teachers and education researchers have long observed that students can appear to reason inconsistently about physical problems.[1] Problems seen as equivalent by experts may not be treated using equivalent reasoning by students. Qualitative research (based on interviews and analysis of open-ended problem solving) has documented many different clusters of semi-consistent reasoning students use in responding to physics problems. This knowledge has been used in creating attractive distracters for multiple-choice examinations[2] that allow one to examine large populations.[3]

The way that students select wrong answers on such tests contains a large amount of valuable information on student understanding. Traditional analyses of multiple-choice exams focus on the scores – the fraction of students that answer each question correctly, and possibly on the correlation between correct answers chosen by students. Such an analysis often fails to explain how students produce incorrect answers. Based on the understanding of student learning developed from qualitative research, we have developed algorithms to conveniently extract and display such information.[4]

The basic idea of our method is to consider that a student's long term knowledge is organized into productive context-dependent patterns of association we refer to as schemas. As a result of different judgments about context made by students and experts, students can appear to experts to function as if they have multiple (possibly contradictory) schemas at the same time. Our method is particularly useful when a population of students responds to a class of physics situations with a small number of fairly robust schemas. This circumstance has been demonstrated by physics education research to be fairly common over a wide variety of physics topics and populations.

Our method allows us to analyze the complete student responses rather than just identifying the fraction of the time they are using the correct approach. The information obtained will be useful only if the test is carefully designed with a good understanding of the student schemas involved with each concept. In this paper we discuss an analytical method for analyzing the concentration /

diversity of student responses to particular multiple-choice questions. This method is both a tool to extract information from a research-based multiple-choice test and a tool to be used in the cyclic process of creating such a test. A method for evaluating and describing the mixed mental state of a class will be described in later papers.[5]

We begin the paper in section II by giving a brief overview of the theoretical structure we use to describe student knowledge. In section III we define the concentration factor, a function that maps the response of a class on a multiple-choice question to the interval [0,1] with zero corresponding to students selecting a random distribution of answers and one corresponding to all students selecting the same answer. In section IV we demonstrate how one can use the concentration factor for all and for incorrect answers to analyze a multiple-choice test. In section V we apply this analysis as an example to the FCI, using data from 14 classes of introductory calculus-based physics for engineers at the University of Maryland ($N = 778$). In section VI we discuss how a concentration analysis can be used in designing and developing a research-based multiple-choice test. We conclude with a summary.

II. A Model of Student Knowledge

We work within a framework developed from what has been learned in neuroscience, cognitive science, and education research. Research in cognitive science and neuroscience has begun to combine to create an understanding of the structure of human memory. Necessarily (and appropriately), most research has been focused on the simplest possible (but still difficult) issues: what is the nature of working memory, how does learning take place in terms of real biological structures, etc. Although researchers have developed a variety of models, there is reasonable agreement on the core elements and structures. In particular, we rely on the following principles:[6]

1. Memory is associative.
2. Cognitive responses are productive.
3. Cognitive responses are context dependent (including the context of the student's state of mind).

To understand the learning of complex subjects, such as college-level physics, we must step beyond models that can currently be confirmed by neuroscience and ask how long-term memory is structured. To understand this, we focus on the following structures that have been proposed by various researchers in neuroscience, cognitive science, and education researchers.[7]

1. Patterns of associations (neural nets)
2. Primitives / Facets
3. Schemas
4. Mental models
5. Physical models

We use these terms in the following way. The pattern of association is the fundamental linking structure represented by connections of neurons and neural net models. An association between elements of memory (declarative or procedural) is context dependent and, since all the factors determining an activation cannot be specified, must be treated probabilistically. Knowledge includes declarative and procedural elements, with procedures being used whenever possible to regenerate recurring patterns as needed in a particular context. A *primitive* is a rule, often indivisible to the user, that when applied in a physical context, produces a *facet* – a statement about how a particular physical system behaves. Declarative knowledge, primitives, and facets are linked in associative patterns that are context dependent. When a particular pattern (containing few or many elements) is robust and occurs with a high probability in particular contexts, we refer to the pattern of association as a *schema*. We call schemas that are particularly robust and coherent *mental models*. If a mental model is based on a set of ideas about physical objects and their properties we call it a *physical model*.

We assume that we are considering a physics topic that has been well-studied using qualitative research methods and that a small number of common naïve schemas or mental models have been identified. We now turn to the question of how to determine the effectiveness of a particular multiple-choice question in triggering this variety of mental models in a population.

III. The Concentration Factor

As we learn from qualitative research into student learning, student responses to problems in many physical contexts can be considered as the result of their applying a small number of mental models. If a multiple-choice question is designed with these alternatives included as distracters, student responses should be concentrated on the choices associated with those models. On the other hand, if the students have little knowledge of the subject, they may act as if they have no models at all, or as if they choose from a wide variety of different models. In this case, their responses will be close to a random distribution among all the choices. Therefore, the way in which the students' responses are distributed can yield information on the students' state.

Type	A	B	C	D	E
I	20	20	20	20	20
II	50	10	30	5	5
III	100	0	0	0	0

Table 1. The possible distribution patterns of student responses when giving a 5-choice multiple-choice question to 100 students.

Choosing a Concentration Factor

Suppose we give a multiple-choice single-response (MCSR) question with 5 choices (A, B, …, E) to 100 students. Some possible distributions of the responses for this question are given in table 1. The types of distributions shown there represent different concentrations of student solutions.

Type-I represents an extreme case where the responses are evenly distributed among all the choices, just like the results of random guessing. Type-II is a more typical distribution that may occur in our classes; there is a higher concentration on some choices than on others. Type-III is the other extreme case where every student has selected the same choice, giving a 100% concentration.

It is convenient to construct a simple measure that gives the information on the distribution of the responses. We define the *concentration factor,* ***C*****,** as a function of student response that takes a value in [0,1]. Larger values represent more concentrated responses with 1 being a perfectly correlated (type III) response and 0 a random (type I) response. We want all other situations to generate values between 0 and 1.

To construct this measure, suppose we give a single MCSR question with m different choices to N students. A single student's response on one question can be represented with a m-dimensional vector $\vec{R}_k = (y_{k1}, \ldots\, y_{ki}, \ldots , y_{km})$, where $k = 1, \ldots, N$ represents different students and $y_{ki} = 1$ (0) if the i^{th} choice is selected (not selected). With a MCSR question, only a single component of $\vec{R}_k$ is non-zero and equals 1. By summing the $\vec{R}_k$ on one question over students we get the total response vector for the question:

$$\vec{R} = \sum_{k=1}^{N} \vec{R}_k = (n_1, n_2, \ldots, n_i, \ldots, n_m) \tag{1}$$

where n_i is the total number of students who selected choice i. Since there is a total of N responses, we have

$$\sum_{i}^{m} n_i = N \tag{2}$$

We can see that the length of $\vec{R}$ actually provides the information on the concentration. For a type-III response (see table 1)

$$\left|\vec{R}\right| = N \tag{3}$$

and for a type-I response

$$\left|\vec{R}\right| = \sqrt{\left(\frac{N}{m}\right)^2 \times m} = \frac{N}{\sqrt{m}} \tag{4}$$

We demonstrate below that all the other situations generate values between $N\big/\sqrt{m}$ and N. Given this circumstance, we can easily construct a concentration measure by subtracting the minimum length and renormalizing. Define r as the scaled length of $\vec{R}$. We can write

$$r = \frac{\sqrt{\sum_{i=1}^{\mathrm{m}} \mathrm{n}_{\mathrm{i}}^2}}{N} \tag{5}$$

where

$$\frac{1}{\sqrt{m}} \le r \le 1 \tag{6}$$

We choose $\boldsymbol{C}$ by subtracting the minimum length from r and renormalizing:

$$\boldsymbol{C} = \frac{\sqrt{m}}{\sqrt{m}-1} \times \left(r - \frac{1}{\sqrt{m}}\right) = \frac{\sqrt{m}}{\sqrt{m}-1} \times \left(\frac{\sqrt{\sum_{i=1}^{m} n_i^2}}{N} - \frac{1}{\sqrt{m}}\right) \tag{7}$$

As a simple check, it is easy to see that when one of the n_i's, e.g. n_j, equals N (and the rest equal 0), $\boldsymbol{C}$ is equal to 1. If all the n_i's are equal (= N/m), $\boldsymbol{C}$ becomes zero.

Finding the Minimum Value of C

To show that all other cases generate values between 0 and 1, we prove that ***C*** has only one minimum equal to zero at $n_i = N/m$. To do this, we can use the Lagrange multiplier method. This problem is equivalent to finding the minimum value of $\left|\vec{R}\right|^2$ under the constraint of Eq. (2). Thus we can write:

$$s = \sum_{i=1}^{m} n_i^2 - \lambda\left(\sum_{i=1}^{m} n_i - N\right) \tag{8}$$

where λ is the LaGrange multiplier. The extreme of $\left|\vec{R}\right|^2$ occurs at $\nabla s = 0$ with λ chosen to satisfy the constraint. To find this extreme point we can do the following:

$$\frac{\partial s}{\partial n_j} = 2n_j - \lambda = 0$$

$$n_j = \frac{l}{2} \tag{9}$$

Since *j* is arbitrary, we have $n_1 = \ldots = n_m = \frac{\lambda}{2}$ and the constraint implies $\sum_{i=1}^{m} n_i = m\frac{\lambda}{2} = N$, which yields

$$\lambda = \frac{2N}{m} \tag{10}$$

At this extreme point, $\left|\vec{R}\right|^2$ can be calculated as:

$$\left|\vec{R}\right|^2_{extreme} = \sum_{i=1}^{m} n_i^2 = m\left(\frac{\lambda}{2}\right)^2 = m\left(\frac{N}{m}\right)^2 = \frac{N^2}{m} \tag{11}$$

Because the largest value of $\left|\vec{R}\right|$ is equal to *N*, it is obvious that this extreme is not a maximum. The second derivative of $\left|\vec{R}\right|^2$ is

$$\frac{\partial^2 \left|\vec{R}\right|^2}{\partial n_j^{\,2}} = 2 > 0 \tag{12}$$

Therefore, this extreme must represent a minimum.

IV. Concentration Analysis

In the following sections, we introduce several methods of using the concentration factor to study different aspects of the student data.

Classifying the Response Patterns

The first method is to combine the concentration factor with scores to form response patterns. The simplest way is to use a two-level coding to characterize the student scores and the concentration factor. For example a question with low score but high concentration will be denoted as an LH type response. The response patterns not only provide a measure of students' performance but also indicate whether the question triggers a common "misconception". Furthermore, the pattern of the shift from pre- to post-instruction tells how the "state" of a class evolves with instruction. For example, the type LL often indicates that most of the students have no dominating model on the topic (as revealed by the test being used) and their responses are close to the results of random guesses. On the other hand, with similar scores, the type LH implies that the test triggers a strong incorrect model on the concept. The response types will not give the detail of the student models but it can show if the questions trigger some common "misconceptions".

In our analysis, we choose a 3-level coding system with "L" for low, "M" for medium and "H" for high. To develop an appropriate quantization scheme, we did simulations for a five-choice test with 100 student responses (m = 5, N = 100). Based on the calculations,[8] we decided to choose a 3-level coding scheme as defined in table 2.

Score (S)	Level	Concentration (***C***)	Level
0~0.4	L	0~0.2	L
0.4~0.7	M	0.2~0.5	M
0.7~1.0	H	0.5~1.0	H

Table 2. Three-level coding scheme for score and concentration factor

A typical research-based MCSR test like the FCI usually has one correct answer and one or more distracters. If the students get low scores, their responses are typically either evenly distributed among the different distracters or concentrated on one or two of the distracters. Combining the ***C*** factor with scores, we can display the different types of responses. We describe them using the following categories (also see table 3):

One-Peak: Most of the responses are concentrated on one choice (not necessarily a correct one).

*Two-Peak :*Most of the responses are concentrated on two choices, usually one correct and one incorrect.\

Non-Peak: The responses are somewhat evenly distributed among three or more choices.

		Implications of the patterns
One-Peak	HH	One correct model
	LH	One dominant incorrect model
Two-Peak	LM	Two possible incorrect models
	MM	Two popular models (correct and incorrect)
Non-Peak	LL	Near random situation

Table 3. Combining score and concentration factor, we can code the student response on a single question with a response pattern. This table shows typical response patterns when using the three-level coding system.

The "One-Peak" situation is typical for either an LH or an HH type of response. In an LH case, students have low scores and most of them picked the same distracter. Therefore it could be considered as a strong indication that the question triggers a common incorrect student model.

The "Two-Peak" situation happens when many of the responses are concentrated on two choices. If one of the two is the correct answer, the response type is an MM; if both choices are incorrect, the response type will be an LM. This type of response indicates that a significant number of students use one or two incorrect models depending on the structure of the questions. Sometimes two incorrect responses can be the result of a single incorrect model.

The "Non-Peak" situation happens when student responses are somewhat evenly distributed over three or more of the choices. The response pattern is usually an LL. This implies that most of the students don't have a strong preference for any models on this topic and the responses are close to the results of random guesses.[9]

Graphical Representation: The S-C Plot

With information on both score and the ***C*** factor, we can construct an "S–***C***" plot, using the score as the abscissa and the concentration as the ordinate. Then the students response on each question can be represented as a point on the S–***C*** plot. Due to the constraint (eq. (2)) there is an entanglement between the score and the concentration factor. As a result, data points can only exist in certain regions on an S–***C*** plot. The boundary of this allowed region can be found mathematically:

Consider the case where we have responses from 100 students with a 5-choice MCSR question (*N* = 100, *m* = 5). Denote the score with *S*. We then have (*N*–*S*) responses left to be distributed among the remaining 4 choices. The smallest ***C*** we can get is when all the (*N*–*S*) responses are closest to an even distribution among the 4 choices. The largest ***C*** occurs when all the (*N*–*S*) responses are concentrated on one of the 4 choices. Therefore we can write

$$\boldsymbol{C}_{MIN}(S) = \frac{\sqrt{5}}{\sqrt{5}-1} \times \left(\frac{\sqrt{4\left(\frac{N-S}{4}\right)^2 + S^2}}{N} - \frac{1}{\sqrt{5}} \right) \tag{13}$$

and

$$\boldsymbol{C}_{MAX}(S) = \frac{\sqrt{5}}{\sqrt{5}-1} \times \left(\frac{\sqrt{(N-S)^2 + S^2}}{N} - \frac{1}{\sqrt{5}} \right) \tag{14}$$

Using eqs. (13) and (14), the boundary of the allowed region is plotted in figure 1. The regions for the six response types are also marked out based on the 3-level quantization scheme in table 2.

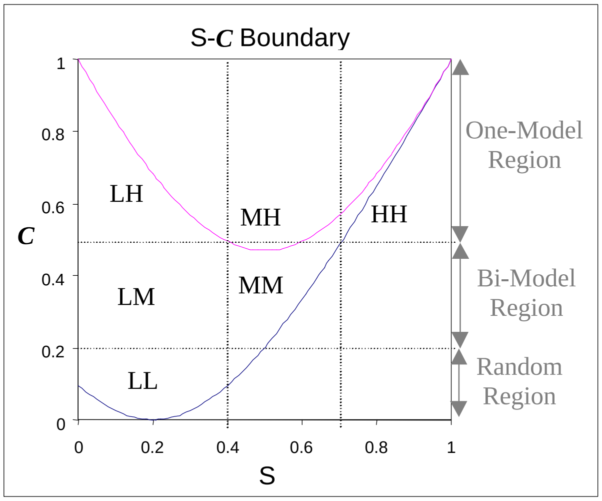


*Figure 1. Combining score and concentration factor, we can create an S-**C** plot to show the score and concentration results of individual multiple choice questions. Due to the constraint between the score and concentration factor, data points can only exist in the area between the two boundary lines.*

In figure 1, the three different situations of concentration – L (no peak), M (two peak), and H (one peak) are also associated with three different indications of possible student model conditions: Random Region –no dominant models; Bi-model Region – two possible models; One-model Region – one dominant model.

Since the number of students is usually very large, the number for all the possible combinations of the students' responses is huge. Defining each possible combination as a state, we can simulate the attractor for random responses by assuming all the responses generated by students are based on random guessing. Figure 2 is a computer simulation of the random attractor obtained with 5 million runs. The value of the density is logarithmic so that we can see more details of the low-density area. As expected, the attractor (the dark area) is concentrated around the minimum point (*S* = 20, ***C*** = 0) with Δ*S* = ±10% and Δ***C*** = 10%. According to our 3-level quantization scheme, this random region is at the center of the LL zone.

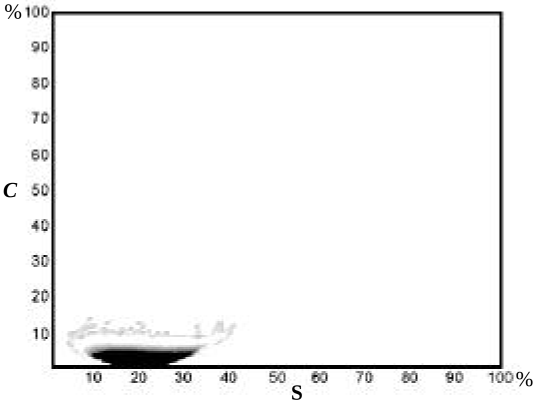


*Figure 2. Assuming all students in a class are guessing, we can simulate the possible structures of student answers on a single multiple-choice question. This generates the S-**C** random attractor. The darker color represents higher probability. At the boundary of the attractor, gray dithering is used to illustrate the boundary line.*

Concentration of the Incorrect Responses: The S-G Plot

The concentration factor gives the overall structure of student responses and is dependent on the score. When the score is high, students necessarily have chosen a single dominant response, so ***C*** will have to be close to 1. In order to disentangle the concentration and the score and to see more detail of the distribution of the incorrect responses, we can define a new concentration variable. From eqs. (13) and (14) it is easy to see that the score determines the absolute boundary of the concentration. The variation of ***C*** within the boundary at a certain score is determined by the distribution of the incorrect student responses. Therefore, if the detail of the distribution of the incorrect responses is of the interest, we need to remove the absolute offset created by the score. This can be done by calculating the concentration for the incorrect responses. Define this as the *concentration deviation*, Γ analogously to ***C*** by

$$\Gamma = \frac{\sqrt{m-1}}{\sqrt{m-1}-1} \times \left(\frac{\sqrt{\sum_{i=1}^{m} n_i^2 - S^2}}{(N-S)} - \frac{1}{\sqrt{m-1}} \right) \qquad (15)$$

Eq. (15) is intrinsically similar to eq. (7) except that the score (correct response) is removed from the sum. This makes Γ and *S* independent. Whatever the score, Γ can have any value within the full range of [0, 1]. We can also construct an S-Γ plot to study the details of the incorrect responses. Since we now have two independent variables as the axes, there is no restriction on the plotting area.

Although Γ has the advantage of being independent of the score and it also provides direct information on the incorrect responses, the measure of the total concentration is still important especially when evaluating the overall model condition. Therefore in order to properly model the student responses, we often need to consider both ***C*** and Γ for different aspects of the data.[10]

Types	LL	LM	LH	ML
Questions	15, 24	5, 9, 18, 28	2, 13, 22	3, 7, 21, 26
Types	MM	MH	HH	
Questions	6, 8, 11, 14, 17, 20, 23, 25	12, 16, 29	1, 4, 10, 19, 27	

Table 4. Using the three-level coding scheme, we combined the pre-instruction FCI data from both tutorial and traditional classes (778 UMd students) and identified the responses types. This table shows the different categories of the student pre-instruction response types.

V. Concentration Analysis of FCI Data

As an example of the kind of information a concentration analysis can give about an exam and a population, we apply our method to results taken with FCI pre- and post-tests. The data is taken from 14 classes in the introductory semester of a calculus-based physics course at the University of Maryland.[11] The students are mostly engineering majors. Half of the classes were taught with University of Washington-style tutorials and the other half of classes were using traditional instruction.[12]

The Initial State of Our Population

The pre-instruction FCI data of all 14 classes were analyzed with the 3-level modeling schemes described in table 2. The results are very similar for all classes;[13] therefore, the results of the pre-data analysis are combined.

Table 4 is a list of the pre-test response types for all 29 questions on the FCI test. To avoid bias generated by variations (e.g. sizes) of the individual classes, the results were obtained by combining the student data from all the classes rather than averaging the results of individual classes. The total number of students in this sample is *N* = 778.

As shown in table 5, the student responses can be grouped into seven categories. The HH and MH types show that the students are doing well on those topics even before instruction. The MM type implies that some students are doing well but a significant number of students, usually more than 30%, have a tendency to use a common incorrect model. More interesting results come from the LM and the LH types, which are strong indications for the existence of common incorrect models. The content of the questions suggests that most of the questions with LM and LH types deal with two physics concepts, the Force-Motion relation and Newton III. “Force-motion” refers to the common naive model that assumes that motion requires and unbalanced force, while “Newton III” refers to the common naïve model that assumes the larger or more active agent will produce the larger force.[14] Table 6 shows the percentage of students selecting the most popular distracters of the questions with LH and LM types of responses. A brief consideration of the distracters in the test (original version)[15] confirms that these questions are associated with two naïve mo dels: Force-Motion and Newton III.With low scores and also low concentration (LL type), questions 15 and 24 represent a different situation where the students did not predominantly favor one or two particular choices. Interestingly, both of the questions deal with detailed physical processes that require an integration of various pieces of physics knowledge. To further clarify the exact reason for the distributions in student responses, we need to look at the content of the questions and conduct detailed research. Sometimes, an LL type can be produced by a question with inappropriate representations or by one that misses including what the students really think.

Force and Motion			Newton’s Third Law		
Choice	%	Type	Choice	%	Type

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*	*Q12*	*Q13*	*Q14*	*Q15*
S	0.79	0.33	0.42	0.74	0.25	0.58	0.46	0.60	0.27	0.80	0.45	0.70	0.22	0.63	0.34
C	0.64	0.50	0.17	0.55	0.40	0.34	0.19	0.35	0.23	0.66	0.33	0.51	0.50	0.43	0.11
	HH	LH	ML	HH	LM	MM	ML	MM	LM	HH	MM	MH	LH	MM	LL
	Q16	*Q17*	*Q18*	*Q19*	*Q20*	*Q21*	*Q22*	*Q23*	*Q24*	*Q25*	*Q26*	*Q27*	*Q28*	*Q29*	
S	0.65	0.63	0.23	0.82	0.49	0.47	0.24	0.58	0.34	0.49	0.48	0.77	0.27	0.67	
C	0.50	0.47	0.41	0.70	0.23	0.20	0.50	0.34	0.08	0.24	0.19	0.61	0.28	0.50	
	MH	MM	LM	HH	MM	ML	LH	MM	LL	MM	ML	HH	LM	MH	

Table 5. With UMd students, we calculated the score and concentration values for all 29 FCI questions with pre and post data from both tutorial and traditional classes.

5-c	58%	LM	2-a	66%	LH
9-c	45%	LM	11-d	43%	MM
18-a	63%	LM	13-c	68%	LH
22-c	66%	LH			
28-d	51%	LM			

Table 6. Student pre-instruction responses on FCI questions related to the concept of Force-Motion and Newton III (UMd students with data from both tutorial and traditional classes combined).

Analyzing the S-C Plot

We can use the S- ***C*** plot to visually study the results. The initial states, final states, and the shifts can be represented with points and vectors on the S-***C*** plot, where each point on the graph represents the average result on one question from all students. Since the tutorial and traditional classes have very different shift vectors, the results from the two types of classes are presented separately. Figure 3 gives the S-***C*** plots of pre and post data for both the tutorial and traditional classes. Each point represents a question and the vectors represent the shifts of pre and post results averaging all 29 FCI questions.

It is easy to see that the pre-states for both classes are similar, but the tutorial class has a much larger shift vector towards the direction of higher score with larger concentration, which indicates that more students favor the correct models. From figure 3, we also see that very few of the FCI questions lie near the "high probability" region of the S–***C*** plot shown in figure 2 that corresponds to students guessing randomly, and that many questions have LH and LM types of responses on pre-test. This implies that the FCI has been successful in finding attractive distracters.

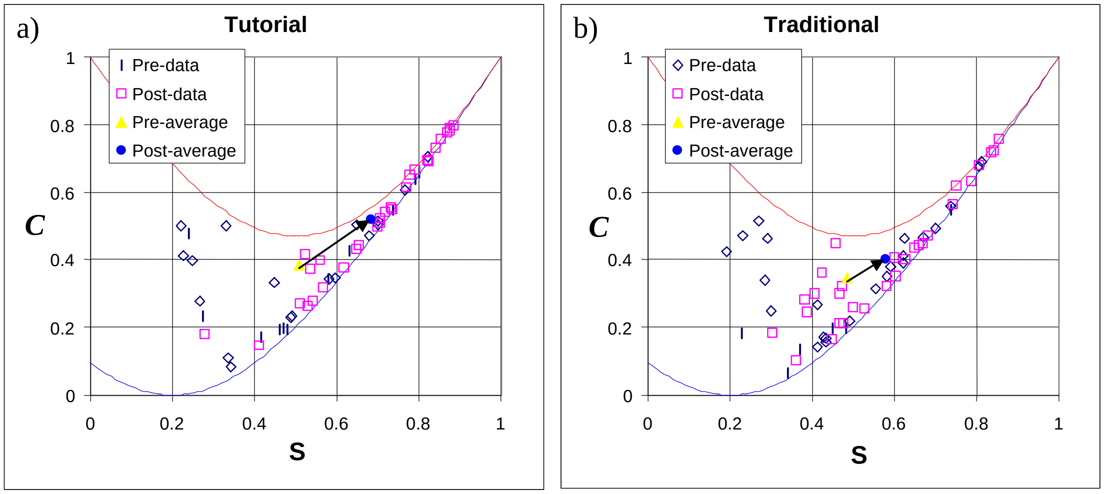


Figure 3. This figure shows S-C plot for all 29 FCI questions with pre and post data from both tutorial and traditional classes. (UMd students).

From table 5, the 29 questions can also be separated into three groups based on student performance measured with pre-instruction scores – high, medium, and low.[16] Since the high performance group is very close to the favorable situation, the low performance group often has much larger contribution on the overall improvement. Therefore the shift of the low performance group should reveal more information about the differences between the two treatments.

The low performance group consists of nine questions with LL, LM and LH types of responses. In figure 4, we plot the S-***C*** shift of these nine questions. The tutorial classes shift towards higher scores and concentrations and the final states are mostly in the HH region. On the other hand, students in traditional classes have some improvement with their scores and the final states are mostly in MM region indicating that a significant number of students still hold an incorrect model and may be in mixed model states.[17]

We can also study the details of student behavior in different concept groups. In figure 5, the shift of the questions in Force-Motion group is plotted. As we can see, the students behave similarly as in the low performance group except that the initial states are mostly in the LM and LH regions indicating a strong initial "misconception". Again, after instruction, the tutorial classes had a large shift bringing the group average close to the HH region. The traditional classes only move to the bi-model region ("two-peak" situation).

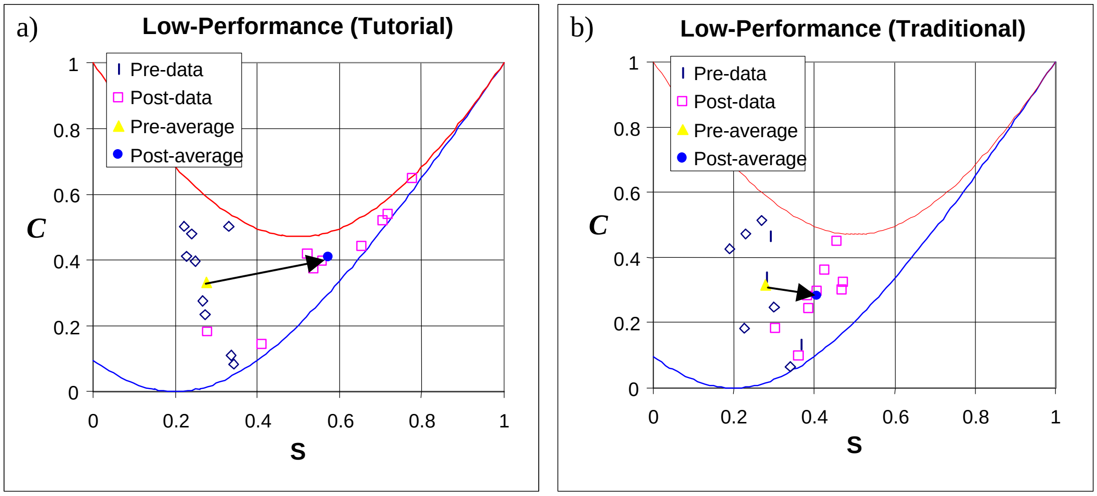


*Figure 4. S-**C** plot for 9 FCI questions (2, 5, 9, 13, 15, 18, 22, 24, 28) with low average pretest scores (<40%).*

When working with the data, we often need to group all the student data together before calculating the average score and concentration. Averaging over results for individual classes (with different sizes) can be mislead and can even yield results outside the allowed region. For example, averaging the two points (0,1) and (1,1) (LH and HH) gives (0.5,1).

Analyzing the S-G Plot

We can also use Γ to study the concentration of the incorrect responses. The average results of Γ for different performance groups is calculated and listed in table 7. We also graph the S-Γ plot for all 29 FCI questions with pre and post data in figure 6. From the data, we can see the interesting

	Tutorial								Traditional							
	Overall		Low		Mid		High		Overall		Low		Mid		High	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
S	0.51	0.69	0.28	0.57	0.55	0.69	0.77	0.83	0.49	0.58	0.28	0.41	0.51	0.60	0.74	0.78
Γ	0.38	0.38	0.53	0.50	0.29	0.31	0.34	0.36	0.35	0.36	0.49	0.50	0.29	0.26	0.35	0.31

Table 7. In this table, we calculated the average values of score and G for FCI questions in different performance groups defined based on pretest scores.

result that the Γ's on low performance questions are consistently higher than that of mid and high performance questions independent of the types of instructions and if the data is taken before or after instruction. Since high Γ's indicate strong distracters, it can be inferred that the low performance questions on the FCI are dominated by situations where student responses have strong alternative models; after instruction the students giving incorrect responses are still strongly affected by certain distracters of these questions.

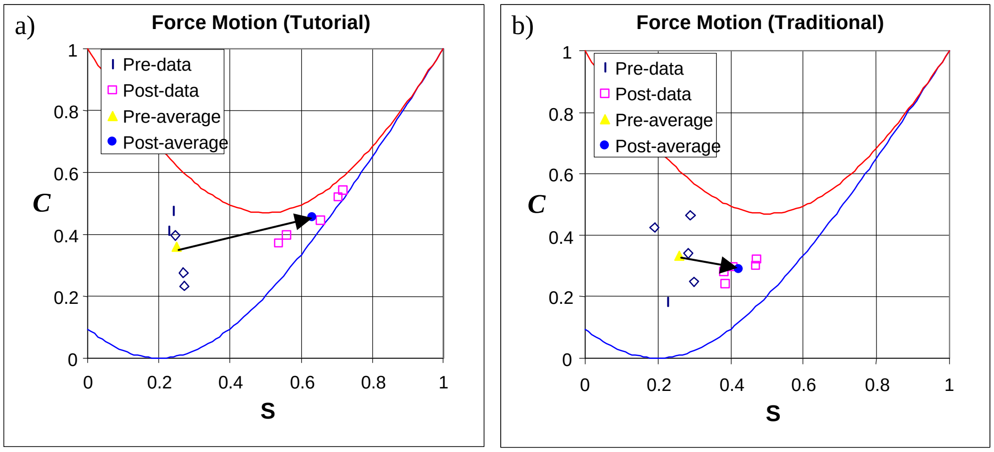


*Figure 5. S-**C** plot of 5 FCI question (5, 9, 18, 22, 28) related to the Force-Motion mental model.*

One advantage of the S-Γ plot is that Γ is not affected by score. From figure 4, the concentration of student post-instructional data gets much larger contribution from the scores and does not show much additional information. On the other hand, even with high scores, the student post Γ's are not affected by scores and are quite scattered just as the results from pre-instruction data (see figure 6). This implies that the students giving incorrect responses behave rather similarly before and after instruction (the students may not be the same). Therefore, using S-Γ plot, we can get more information on students' giving incorrect answers than what can be obtained with S-***C*** plot.

In figure 7, the results of the low performance questions are plotted with the shift vectors for all the questions displayed. This figure dramatically demonstrates that questions on the FCI on which students perform poorly are primarily questions on which our student population holds common alternative mental models. The poor performance does not result from random guessing. (The distribution associated with random guessing can be inferred from the random S-***C*** attractor shown in figure 2.) We can also easily see that not only the average results, but also the shifts of individual questions in the low performance group are similar except for questions 9 and 22. To under-

Question	9						22					
Choice	a	b	*c*	**d**	e	**Γ**	a	b	*c*	**d**	e	**Γ**
Tutorial												
Pre	0.07	0.17	*0.47*	**0.27**	0.01	**0.40**	0.03	0.06	*0.67*	**0.24**	0.01	**0.76**
Post	0.05	0.05	*0.2*	**0.70**	0.00	**0.42**	0.10	0.03	*0.2*	**0.66**	0.02	**0.30**
Traditional												
Pre	0.06	0.27	*0.42*	**0.24**	0.02	**0.29**	0.01	0.05	*0.66*	**0.27**	0.01	**0.81**
Post	0.04	0.08	*0.38*	**0.49**	0.01	**0.55**	0.07	0.10	*0.42*	**0.40**	0.01	**0.51**

Table 8. Student responses on FCI questions 9 and 22 where the correct choice is shown in bold and the major distracter is italicized.

stand this phenomenon, we first analyze these two questions in details. The student responses on FCI question 9 and 22 are listed in table 8.

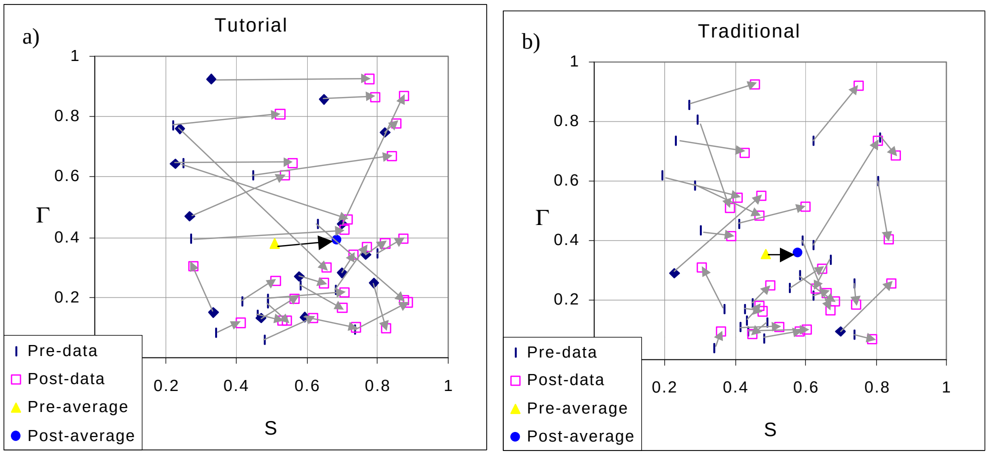


Figure 6. S-G plot for all 29 FCI questions with pre and post data from both tutorial and traditional classes. (UMd students).

As we can see, for FCI question 9 (shown in figure 8), the incorrect responses of the students in tutorial classes are all significantly reduced after instruction. This results in a Γ similar to that of the pre-instruction data. On the other hand, the incorrect responses of students with traditional instruction only have minor changes except for a large drop on choice "b". Therefore the post-data has a very high Γ with student responses concentrating on the main distracter (choice "c"). The only difference between choice "b" and "c" is that in choice "c" a "normal force" is included (both "b" and "c" follow the belief that there is a force in the direction of motion). This result indicates that after traditional instruction students are much improved on recognizing the "normal force", however, many of them still hold their initial belief that a force is needed in the direction of motion.

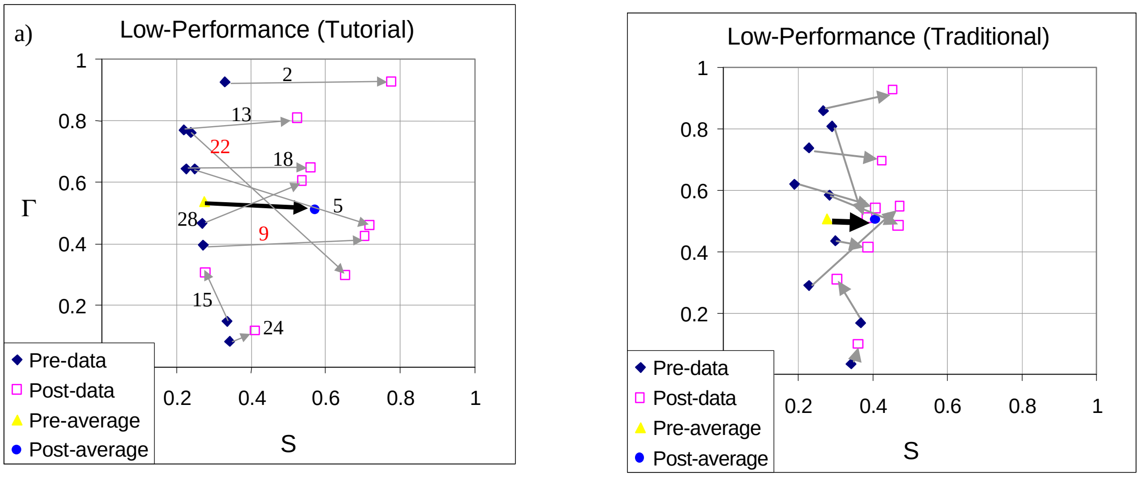


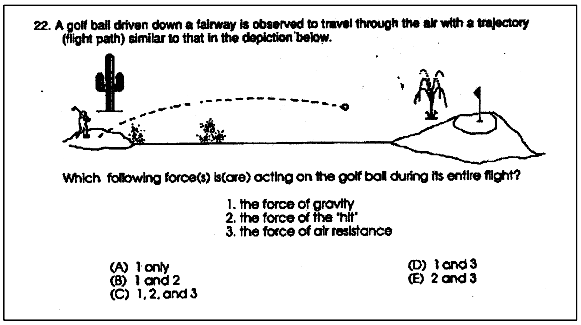
Figure 7. S-G plot for 9 FCI questions (2, 5, 9, 13, 15, 18, 22, 24, 28) with low average pretest scores (<40%)

For FCI question 22 (shown in figure 9), the data shows only one major distracter (choice "b"). The variations of student responses on other distracters are around 5%. Therefore in this question, Γ depends mostly on the student response on the main distracter. When students get large improvement, as it is in the tutorial classes, the post-Γ is significantly lower than the pre-Γ. Students with traditional instruction have much less improvement and the post-Γ is still quite high.

On other questions, the pre and post Γ's have similar values. In tutorial classes, student improvement on scores is comparatively large and the number of student responses on the major distracter is significantly reduced. The similar pre and post Γ's are mainly produced by simultaneous decreases on most incorrect responses. In tradional classes, student improvement on scores is often small, which results in much less impact on Γ's. In general, for traditional classes, the student pre and post results remain similar (~ 15% changes).

VI. Discussion and Summary

The concentration factor can be used in many ways in both research and instruction. In research, we can use it to facilitate the design of effective multiple-choice questions that can be used to probe student conceptual understanding. In instruction, with a research-based multiple-choice test, we can use the concentration factor to evaluate student performance and their modeling conditions.



22. A golf ball driven down a fairway is observed to travel through the air with a trajectory (flight path) similar to that in the depiction below.

Which following force(s) is(are) acting on the golf ball during its entire flight?

1. the force of gravity
2. the force of the "hit"
3. the force of air resistance

(A) 1 only
(B) 1 and 2
(C) 1, 2, and 3
(D) 1 and 3
(E) 2 and 3

Figure 8. FCI question 22

Facilitating Test Development

In PER and education research in other areas, many researchers are working to develop effective multiple-choice tests in order to be able to evaluate and compare instruction that is delivered to large populations. Useful multiple-choice tests may be created in situations in which systematic research on student understandings of the physics concepts has demonstrated the presence of common naïve models in a particular populat ion. In this situation, distracters in multiple-choice questions can be designed to probe the distribution of these models. Once a prototype is proposed, it has to be tested and validated with further research. In this process, the concentration factor can be used to help further the development of the test in two ways.

1. *A concentration analysis can help confirm the presence (and level) of erroneous models detected through research.*

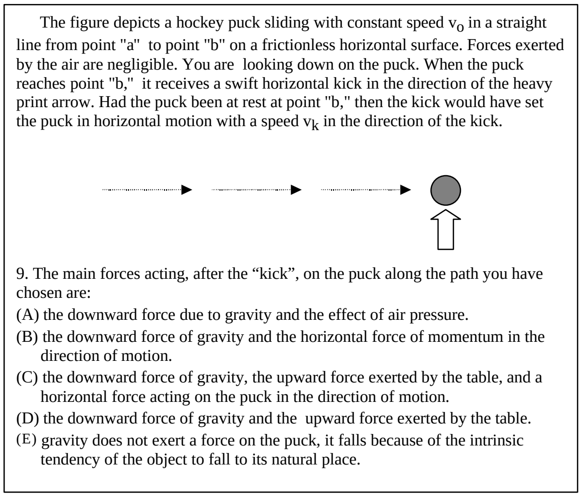
The design of a test usually starts with detailed student interviews where the incorrect student models can be identified. Then we design the multiple-choice questions with distracters associated with these incorrect student models. Using the concentration factor to analyze the results of the test, we can obtain quantitative evaluations and evidence on whether these distracters match well with the student models, and/or if the student models detected in interviews are common to a large population of students.[18] If a distracter is effective, we often observe a low score but high ***C*** and Γ with students before instruction.

2. *A concentration analysis allows one to detect items where a relevant distracter may be missing or existing ones ineffective.*

When a question is designed appropriately, we usually will observe an LH or LM type of response with pre-test data. If the result shows an LL type of response, it indicates that the distrac ters are not attractive. This can be caused by three possible situations: 1. None of the distracters reflects a common student model; 2. For the context of the question, there does not exist a common student model; 3. All the choices correspond well with the student models, and the students are using all the models equally. When this happens, it often indicates that more research is needed to further clarify the details involved.

3. *A concentration analysis can help improve any multiple-choice instrument.*

The concentration factor gives a way to automate the selection of interesting items in any existing test. For example, using an S-Γ plot, we can quickly scan many items and select the ones that might be particularly interesting to look at in detail. Then we can conduct qualitative research on these interesting items to determine if the students have common incorrect models and if the questions are detecting these models, and use the results to redesign the questions. Of course, if the test is to be effective, the first version must be based both on a good understanding of what is to be learned and on sound insights into student thinking, however obtained.



The figure depicts a hockey puck sliding with constant speed v_o in a straight line from point "a" to point "b" on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point "b," it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point "b," then the kick would have set the puck in horizontal motion with a speed v_k in the direction of the kick.

9. The main forces acting, after the "kick", on the puck along the path you have chosen are:

(A) the downward force due to gravity and the effect of air pressure.

(B) the downward force of gravity and the horizontal force of momentum in the direction of motion.

(C) the downward force of gravity, the upward force exerted by the table, and a horizontal force acting on the puck in the direction of motion.

(D) the downward force of gravity and the upward force exerted by the table.

(E) gravity does not exert a force on the puck, it falls because of the intrinsic tendency of the object to fall to its natural place.

Figure 9. FCI question 9

When we study student modeling, the questions should be carefully designed so that the distracters match the common incorrect models. To achieve best results, it is helpful to have a single choice on each question representing one common student model. The number of choices in each question is also an important factor. A small number of choices can generate large distortion on student responses. In addition, with a small number of choices (≤ 3), a multiple-choice question becomes close to a true-or-false question. It is then less meaningful to use the concentration evaluation, since once the score is known, the student incorrect responses are also obvious. We suggest that the number of choices for each question should be no less than 5. This reduces the probability that a student guessing at random will select a choice corresponding to a known model. (See ref. 4 for a more extended discussion of this point.)

Furthermore, to keep consistency in calculating the concentration factor, it is recommended to design the questions so that they all have the same number of choices. However, when the number of choices is large (>6), small variations (±1) on the numbers of choices for different questions often result in differences that can be tolerated.[19]

Facilitating Instruction and Assessment

In instruction, when we have a research-based test available, we can use the concentration factor to evaluate student performance and the effectiveness of instruction. Traditionally, student performance is evaluated with scores, which only gives limited information on student understanding especially with low scores. The information on how the majority of students get a question wrong cannot be reflected using scores alone. This information can be an important clue for instructors to help them improve their teaching.

With the concentration factor, we can retain part of the information on students incorrect answers and infer the states of student mental models. Especially when instruction is integrated with research, we can use concentration factor to evaluate student models on different concepts, and to compare student improvement with different instructional methods.

In this paper, we have introduced a new method to study the structure of the student responses on a multiple-choice test that provides useful information on the distribution of student responses. The results can be used to analyze the conditions of student mental models. Applications with FCI data confirm many widely recognized results and the additional information obtained with this method gives new ways to study the student difficulties. This method can be a useful tool to provide guidance in the development of more effective multiple-choice tests and as a part of comprehensive assessment of a class's learning.

Acknowledgment

We would like to thank Dean Zollman, Alan Van Heuvelen, Richard J. Furnstahl, Mel Sabella, and Michael Wittmann for their many constructive discussions. This work is part of the research supported by NSF grant DUE 965-2877.

Endnotes

[1] I. A. Halloun and D. Hestenes, "Common sense concepts about motion", Am. J. Phys. **53**, 1056 (1985)., McDermott, diSessa, D. P. Maloney and R. S. Siegler, "Conceptual competition in physics learning." *International Journal of Science Education,* **15**, 283-296 (1993); R. K. Thornton, "Conceptual Dynamics: Changing Student Views of Force and Motion," *Proceedings of the International Conference on Thinking Science for Teaching: the Case of Physics*, Rome, Sept. 1994; M. Wittmann, "Making Sense of How Students Come to an Understanding of Physics: An Exam-
esis, University of Maryland, 1999.

[2] I. A. **53**, 1043
1056 (1985).; D. Hestenes, M. Wells and G. Swack **30**, 141 158
(1992). D. Hestenes and M. Wells, "A Mech **30**, 159 166 (1992); R. J. Beichner,
tics graphs," Am. J. Phys. , -762 (1994);
Sokoloff, "Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the
ation of active learning laboratory and lecture curricula," Am. J. Phys. , 338-

[3] -engagement versus traditional methods: A six thousand-
data for introductory physics courses," . **66** -74 (1998).

[4] Lei Bao, "Dynamics of Student Modeling: A Theory, Algorithms, and Application to Quantum Mechanics," Ph.D.

[5] F. Redish, "Diagnosing student problems using the results and methods of physics educa
research," to be published in *Proceedings of the 1999International Conference of Physics eachers and Educators*
held in Guilin, China, Aug. 18- editor

J. M. Fuster, *Memory in the Cerebral Cortex: An empirical approach to neural networks in the human and nonh man primate* (MIT Press, 1999); J. R. Anderson and C. Lebiere, (Erl

1998); T. Shallice and P. Burgess, "The domain of supervisory processes and the temporal organization of beha ior," in *ctions* -35.

[7] Andrea diSessa, “Toward an Epistemology of Physics,” , (1993) 105- nstrell, *Research in Physics Learning: Theoretical Is and* , Bremen, Germany, March 4- Duit, F. Goldberg, and H. Nie derer (IPN, Kiel Germany, 1992) 110- mental models”, in Derdre Gentner and Albert L. Stevens, Eds. *Mental Models* iates, -14; D. E. Rumelhart, “Schema *Comprehension and Teaching: Research Reviews* ociation, 1981) 3 26**.**

[8] reference 4.

Since this is close to the random situation where the e fect of the random variation is large, it will be difficult to diffe entiate whether the individual response is due to systematic reasoning with many different models or guessing. died by qualitative methods e.g. interviews.

[10] M ***C*** and can be found in reference 4.

The topic covered during this semester is Newtonian mechanics. The data was collected by Dr. J. Saul at the Uni-rsity of Maryland (UMd).

[12] L. C. McDermott, P. S. Shaffer, et al., *nt* (Prentice Hall, New York NY, 1998). For details on the application of these tutorials at the University of Maryland, see E. F. Redish, J. M. Saul, and R. N. -engagement m crocomputer- tories,” Am. J. Phys. **65** -54

[13]

[14] ssion.

[15] D. Hestenes, M. Wells, and G. Swackhamer, “Force co cept inventory”, *Phys. Teach.* **30**, 141 151 (1992)

Specifically, the items are classified as follows: Low performance group: 2, 5, 9, 13, 15, 18, 22, 24, 28; Mid per-mance group: 3, 6, 7, 8, 11, 1

[17]

[18] ices. For many que tions, one might want to do this, just as we did for items 9 and 22 of the FCI above. The concentration factor is a way to automate the selection of items to be co sidered. An S- plot, for example, allows one to quickly scan many ite ularly interesting to look at in detail.

[19]