Chapter 3: Model Analysis Algorithms I: The Concentration Factor

Introduction

Students' wrong answers contain a large amount of valuable information. Usually this information is extracted laboriously from analysis of transcripts of interviews and student responses to open-ended exam questions. Another simple and convenient instrument that can be used to study student models is the multiple-choice conceptual test, which can be easily conducted in large scale. There are many examples of such tests: FCI, FMCE, etc. A major disadvantage of the multiple-choice test is that the traditional evaluation method only gives the score and doesn't provide information on what causes the students' problems. Based on our model of student learning of physics, we can now develop algorithms to extract such information. The basic idea is to consider that the student can be in multiple-model state at the same time and to take into account all the student responses rather than just the correct ones. By studying the relations among the different responses, we can extract useful information on the student models that generate these responses.

The information obtained will be useful only if the test is carefully designed with a good understanding of the physical models involved with each concept.[1] We therefore see this method as both a tool to extract information from a research-based multiple-choice test and as a tool to be used in the cyclic process of creating such a test.

In this chapter and chapter 4, I will introduce two algorithms to do quantitative evaluations of student models based on multiple-choice test data. The first algorithm introduces a new evaluation variable – the concentration factor. This measurement can provide quantitative evaluations on the condition of the possible student models. The second algorithm provides quantitative estimations of the actual student models where individual student responses are modeled and stored with a model density matrix. The student model states are then evaluated by analyzing the eigenvectors and eigenvalues of the model density matrix.

Model Condition Evaluation – the Concentration Factor

The multiple-choice test can be a good tool for assessing student understanding, performance, and improvement in a wide range of contexts. A very popular one is the Force Concept Inventory (FCI), which has been widely used by many instructors and physics education researchers.[2] The results are useful in evaluating student understanding of physics as well as investigating the effectiveness of instruction.

Student performance on these tests are usually measured with the number of correct responses. This measurement, although necessary and important, does not give information about the distribution of students' responses. In this chapter, I introduce a new measurement that gives this information. It is called the c*oncentration factor*, which tells about how the student responses are distributed, i.e., if the responses are concentrated on certain choices or widely scattered among the different choices like the results of random guesses.

Based on our understanding of student learning, the student responses are considered as the output of students applying their models in various physical contexts. Therefore if students have a few consistent models of physics, the responses should be more concentrated on those choices representing the corresponding models. On the other hand, if the students have no models, or have a wide variety of models, their responses will be more randomly distributed among the choices. Therefore, the way in which the students' responses are distributed reveals information on the students' modeling conditions.

For best results, it is helpful for the questions to be designed such that the different choices of a question have a one-to-one correspondence to different student models. The interpretation of the concentration can be complicated when a single choice reflects multiple student models or multiple choices in a question are related to one student model. According to our experience with student data on FCI and FMCE test, although questions in these tests are not designed with one-to-one correspondence between student models and the choices, student responses are found to be mostly concentrated on one of the choices corresponding to a student model. If a question is found to have a choice that can be related to multiple student models, it will not be used in our analysis.

The Formulation of the Concentration Factor

Let us look at a simple example. Suppose we give a multiple choice single response (MCSR) question, each with 5 choices labeled A through E, to 100 students. For each question, we will get 100 responses. Table 3-1 lists some possible distributions of the responses for one question.

Table 3-1. Possible distributions of responses from a class of students on one question

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

Type-I represents an extreme case where the response is evenly distributed among the choices, just like the results of random guesses. Type-II is a more typical distribution that may occur in our classes. Type-III is the extreme case where everybody selected the same choice, giving a 100% concentration. We can define Type-III as the one with the highest concentration and Type-I as the lowest.

It will be convenient to construct a simple measure that gives the information on the distribution of the responses. Define it as the *concentration factor*, ***C***, with a value in the range [0,1], where larger values represent more concentrated responses. We would like to have ***C*** equal 1 for a Type-III response, and equal 0 for a Type-I response. All the other types should generate a value between 0 and 1.

Now consider a single MCSR question with "m" different choices and a total of N student responses. A single student's response on one question can be represented with a m-dimensional vector $\vec{r}_k$ = (y_{k1},… y_{ki}, … , y_{km}), where k = 1,…, N represents different students and y_k = 1 (0) if the corresponding choice is selected (not selected). For a single item of an MCSR question, only one component of $\vec{r}_k$ is non-zero and equals 1. Then we can sum up all the student responses on one question in this vector form and get the total response vector for that question:

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

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

$$\sum_{i}^{m} n_i = N \qquad (3\text{-}1)$$

We can see that the length of $\vec{r}$ actually provides the information about the concentration. The value of $|\vec{r}|$ equals N with a Type-III response and it equals $\sqrt{\left(\frac{N}{m}\right)^2 \times m} = \frac{N}{\sqrt{m}}$ with a Type-I response. All the other types will generate a value between $\frac{N}{\sqrt{m}}$ and N. Define r_0 as the scaled length of $\vec{r}$. We can write

$$r_0 = \frac{\sqrt{\sum_{i=1}^{m} n_i^2}}{N}$$

where

$$\frac{1}{\sqrt{m}} \le r_0 \le 1$$

This suggests to choose ***C*** as:

$$\boldsymbol{C} = \frac{\sqrt{m}}{\sqrt{m}-1} \times (r_0 - \frac{1}{\sqrt{m}}) = \frac{\sqrt{m}}{(\sqrt{m}-1)} \times (\frac{\sqrt{\sum_{i=1}^{m} n_i^2}}{N} - \frac{1}{\sqrt{m}}) \qquad (3\text{-}2a)$$

where N >> m is required.

As a simple verification, it is easy to see that when one of the n_i's equals N, which means all the other ones equal 0, ***C*** is equal to 1. If all the n_i's are equal (= N/m), ***C*** becomes zero. All the other cases generate a value between 0 to 1. A 5-choice example can be written as:

$$\boldsymbol{C} = \frac{\sqrt{5}}{\sqrt{5}-1} \times (\frac{\sqrt{n_a^2 + n_b^2 + n_c^2 + n_d^2 + n_e^2}}{N} - \frac{1}{\sqrt{5}})$$

The Derivation

To show that ***C*** has only one minimum equal to zero at $n_i = N/m$, we cam use the LaGrange multiplier method. This problem is equivalent to finding the minimum value of $|\vec{r}|^2$ under the constraint of Eq. (3-1). Thus we can write:

$$|\vec{r}|^2 = \sum_{i=1}^{m} n_i^2 - \lambda\left(\sum_{i=1}^{m} n_i - N\right) \tag{3-3}$$

where λ is the LaGrange multiplier. The extreme of $|\vec{r}|^2$ occurs at $\nabla|\vec{r}|^2 = 0$. To find this extreme point we can do the following:

$$\frac{\partial|\vec{r}|^2}{\partial n_j} = 2n_j - \lambda = 0$$

$$\therefore \quad n_j = \frac{\lambda}{2}$$

Since j is arbitrary, we have $n_1 = \ldots = n_m = \dfrac{\lambda}{2}$ and $\sum_{i=1}^{m} n_i = m\dfrac{\lambda}{2} = N$, which gives

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

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

$$|\vec{r}|^2_{\text{extreme}} = \sum_{i-1}^{m} n_i^2 = m\left(\frac{\lambda}{2}\right)^2 = m\left(\frac{N}{m}\right)^2 = \frac{N^2}{m} \tag{3-4}$$

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

$$\frac{\partial^2|\vec{r}|^2}{\partial n_j^2} = 2 > 0,$$

therefore, this extreme must represent a minimum. Plugging the result of Eq. (3-4) back into Eq. (3-2), we can verify that the minimum of ***C*** is 0 and it happens when all the n_i's are equal.

The Application (How to use this tool)

We can construct different forms of evaluations depending on the information we want. In the following sections, I will introduce several methods of using the concentration factor to study different aspects of the student data.

Using *C* to Evaluate Student Model Condition

- **Response patterns**

The first method is to combine the concentration factor with scores to form response patterns. When comparing the pre and post test results, such pattern shifts can provide more information than just the shift of scores.

The simplest way to consider the patterns of responses is to use a two-level code to label the student scores and the concentration factor. Thus a question with low score but high concentration will be denoted as a LH type response. Table 3-2 is a list of all the possible patterns and shifts in binary coding with their possible indications. An impossible HL type is ignored since high score creates high concentration.

The response patterns not only provide a measure of student performance but also indicate whether the students have dominating misconceptions. Furthermore, the pattern of the shift also tells how the student "states" evolve with instruction.

Table 3-2. The response patterns with two-level coding (Score:Concentration)

Pre Test	Post Test	Indications of the pattern shift
LL	LL	Students haven't learned anything.
	LH	Instruction is leading students to an incorrect model
	HH	Effective instruction in the right direction.
LH	LL	Students had incorrect preconceptions and unlearned.
	LH	Instruction has no effect.
	HH	Instruction made the right shift to the correct concepts.
HH	LL	Complete unlearn.
	LH	Shift towards incorrect models.
	HH	Students are good on this topic.

As a simple example, let us look at the difference between a type LL and a type LH response. The LL type shows that most of the students have no dominating model on the topic (at least as revealed by the test being studied) and their responses are more like the results of random guesses.

On the other hand, with similar scores, the type LH implies that the students probably have a strong incorrect model on the related concept. If the results are from a pre-test, the instructors can be informed by these incorrect initial student models and prepare for appropriate instruction.

When comparing the pre and post-test results, analyzing the shift pattern can be useful. For instance, if we have a LL–LH shift, it indicates that the instruction is making some effects but there exist problems in the learning or teaching that shifts the students towards an incorrect direction. Therefore in such circumstances we need to carefully review the instruction and study the student responses in detail to find out the possible causes of the problems.

In practice the binary coding is a little too coarse. In the analysis, I choose a three-level coding system with "L" for low, "M" for medium and "H" for high. To get an idea of how a typical set of data behaves, I did some simulations as a starting point to develop an appropriate quantization scheme. The calculation is done to simulate a five-choice test with 100 student responses (m = 5, N = 100). The results of three typical types of responses are listed in table B-1 through table B-3 in Appendix B. Based on the calculations, a 3-level coding scheme is define as in table 3-3.

Table 3-3. Three-level coding scheme for score and concentration factor

Score	Level	***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

For a typical MCSR test like FCI, we usually have one correct answer and several distracters. If the students get low scores, their responses could be either evenly distributed among the different distracters or concentrated on one or two of the distracters. Combining the ***C*** factor with scores, we can model the different types of responses. The following is a list of some response situations that we are interested in.

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: Most of the responses are somewhat evenly distributed among three or more choices.

The "One-Peak" type is typical for either a LH or a HH kind of response. In a LH case, students have low scores and most of them picked the same distracter. Therefore it could be considered as a strong indication for a common incorrect student model. As for the HH, it shows that the students are doing well on that topic. It is possible that a high concentration on a response can be a result from a poorly worded or misleading question.

Our discussion assumes the choices have been carefully designed based on systematic research (with detailed individual interviews) and carefully validated so that they represent common misinterpretations. Since both FCI and FMCE tests are developed based on research, our assumption is appropriate for these two tests.

The "Two-Peak" situation happens when many of the responses are concentrated on two of the choices. If one of the two is the correct answer, the response type is a MM; if both choices are incorrect, the response type will be a LM. This type of response indicates that a significant number of students are having one or two misconceptions depending on the structure of the questions. Sometimes two incorrect responses can be the result of the same 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 a LL. This implies that most of the students don't have a strong preference of any models on this topic and the responses are close to the results of random guesses.[3]

- **Graphical representation of the data**

– The S-***C*** plot

With information on both score and the ***C*** factor, the responses and their shift patterns can now be represented on a two dimensional graph. We can construct an "S−***C***" plot, using the score as the horizontal axis and the concentration as the vertical axis. Then the response of each question could be represented as a point on the S−***C*** plot. The shift of the response can be represented with a vector starting from the point representing the initial state towards the point for the final state. Before using this graphical method, it is useful to know some of its basic characteristics.

– The allowed region of S-***C*** plot

Due to the constraint created by the entanglement between score and the concentration, data points may not fall in all the regions on an S−***C*** plot. For example, if we have a response with a score of 100%, the concentration must also be 100%. On the other hand, if the score has some other values, we will get a spread of different possible concentration values. The different combinations of score and concentration can only exist inside a bounded region on the S-***C*** plot. The boundary of this allowed region can be found mathematically as followings:

Consider the case where we have 100 total responses from a 5 choices 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 evenly 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

$$\boldsymbol{C}_{\mathrm{MIN}}(\mathrm{S}) = \frac{\sqrt{5}}{\sqrt{5}-1} \times \left(\frac{\sqrt{4\left(\frac{\mathrm{N}-\mathrm{S}}{4}\right)^2 + \mathrm{S}^2}}{\mathrm{N}} - \frac{1}{\sqrt{5}} \right) \qquad \text{(3-5a)}$$

and

$$\boldsymbol{C}_{\mathrm{MAX}}(\mathrm{S}) = \frac{\sqrt{5}}{\sqrt{5}-1} \times \left(\frac{\sqrt{(\mathrm{N}-\mathrm{S})^2 + \mathrm{S}^2}}{\mathrm{N}} - \frac{1}{\sqrt{5}} \right) \qquad \text{(3-5b)}$$

Using Eq. (3-5a/b), the boundary of the allowed region is plotted in figure 3-1. The regions for the 6 typical response types are also marked out based on the three-level quantization scheme in table 3-3.

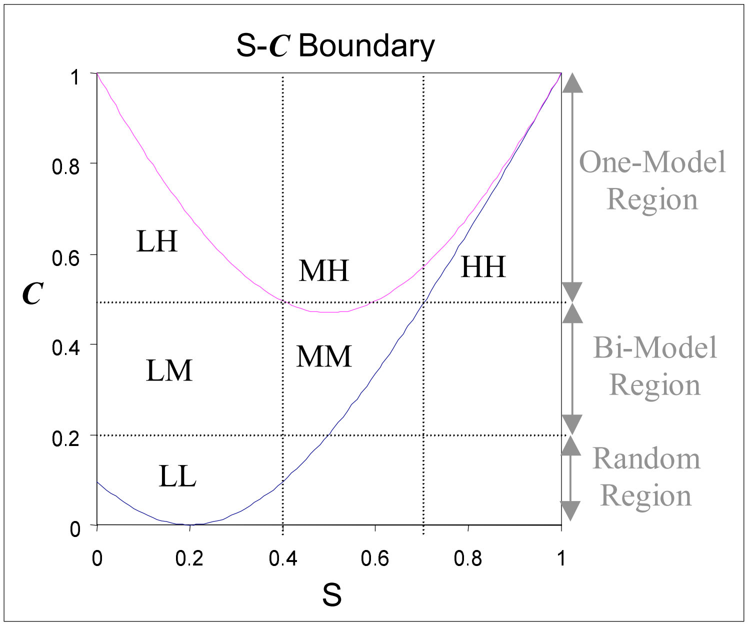


Figure 3-1. Allowed region of the S-***C*** plot (the area between the two boundary lines)

Also shown in figure 3-1, the three different situations of concentration – L (no peak), M (two peak), and H (one peak) are associated with three different indications of possible student model conditions:

1. Random Region: with no dominant models,

2. Bi-model Region: possibly with two popular models,

3. and One-model Region: possibly with one dominant model.

– The state density

Since the total number of responses is usually very large, the number for all the possible combinations of these responses is huge. Again, taking N = 100 and m = 5 for example and defining each possible combination as a state, the total number of all the possible states would be:

$$\mathcal{N} = \sum_{n_1=0}^{100} \left\{ \sum_{n_2=0}^{100-n_1} \left[\sum_{n_3=0}^{100-n_1-n_2} \left(\sum_{n_4}^{100-n_1-n_2-n_3} 1 \right) \right] \right\} = 4598126 \;\cong 4.6 \times 10^7$$

where n_i represents the number of students that select the ith choice.

It is interesting to see how all these states are distributed (at least mathematically). Figure 3-2 shows the computer simulation of the state density by assuming that all states are equally probable to occur.

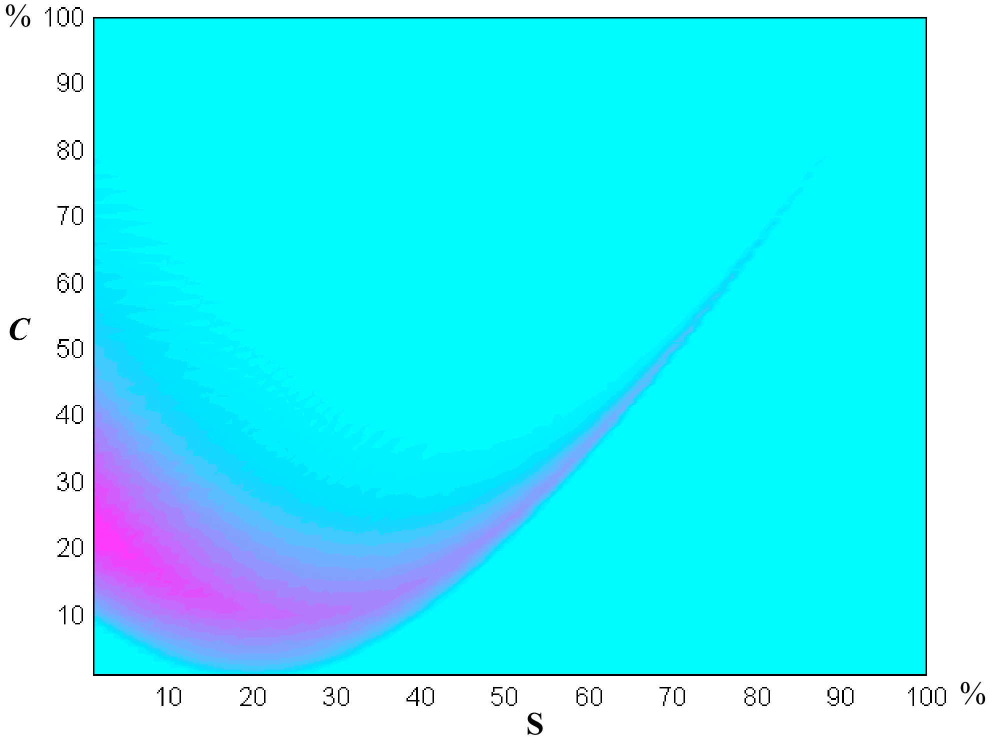


Figure 3-2. S-***C*** state density attractor

The hot (red) colors in the graph represent higher density of states and the cool colors represent lower density. It can be shown that the centerline, representing the maximum density, is the low boundary of the S-***C*** plot under an additional constraint that one of the choices is 0. (See Appendix B for more details.)

– The random response attractor

In reality, the probability for each state is affected by the students and the questions in the test. For a special case where we assume all the responses generated by students are based on random guessing, it is possible to simulate the attractor for random responses. Figure 3-3 is a computer simulation of the random attractor.

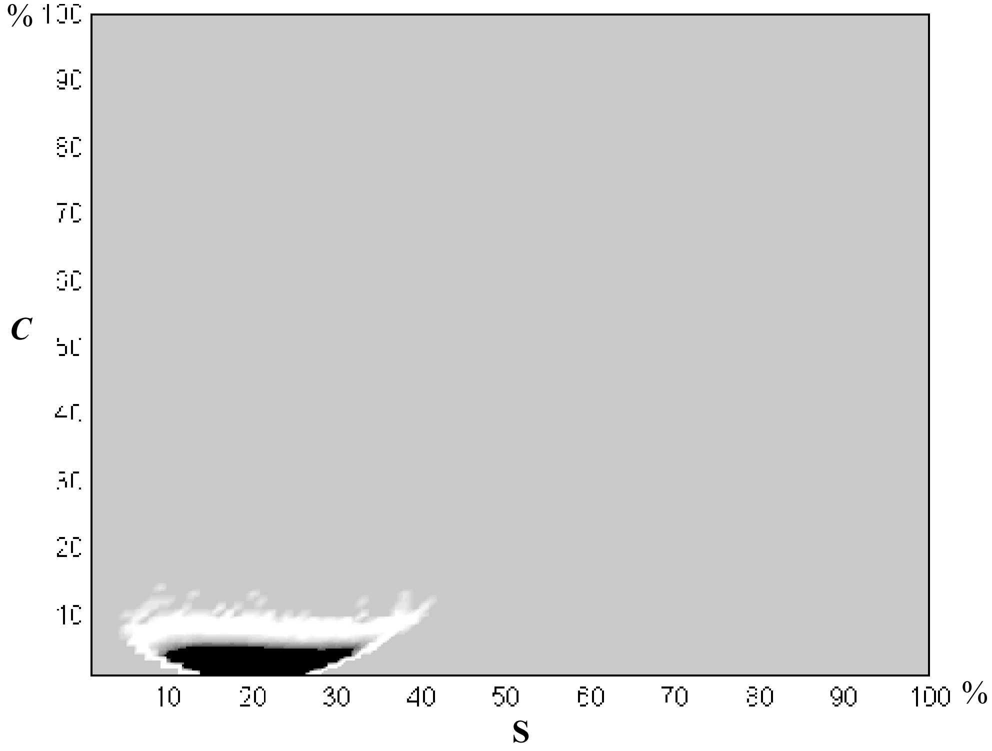


Figure 3-3. S-***C*** random attractor

The simulation is done with a computer that is programmed to generate huge number of random responses. Figure 3-3 represents an attractor of 5 million runs. The value of the density is logarithmic so that we can see more of the low-density area. As expected, the attractor (the dark concentration) is concentrated around the minimum point (S = 20, ***C*** = 0) with ΔS = ±10% and Δ***C*** = 10%. According to our three-level quantization scheme, this random region is at the center of the LL zone.

Concentration of the Incorrect Responses

The S-***C*** plot represents the overall concentration of student responses. Due to the constraint of Eq. (3-5), the data points on an S-***C*** plot are restricted in a strangely shaped attractor. In addition, the ***C*** given by Eq. (3-2) is dependent on the score. Such dependence will strongly affect the meaning of the overall concentration especially at large scores where most of the contributions are generated by the scores (the information on the incorrect responses are overwhelmed by the scores). In order to disentangle the concentration and the score, we design a new variable. From Eq. (3-5) it is easy to see that the score determines the absolute boundary of the concentration. The variation of ***C*** within the boundary 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. Defining it as the *concentration deviation*, Γ, we can write

$$\Gamma = \frac{\sqrt{m-1}}{\sqrt{m-1}-1} \times \left(\frac{\sqrt{\sum_{i=1}^{m} n_i^2 - S^2}}{(N-S)} - \frac{1}{\sqrt{m-1}} \right) \qquad (3\text{-}10)$$

Eq. (3-10) is intrinsically similar to Eq. (3-2) except that the score (correct response) is removed. This makes Γ and S independent and Γ can have any value within the full range of [0, 1] no matter what value S has. We can 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.

Because Γ only deals with the incorrect responses, in many cases, there is only one concentration peak on one of the incorrect choices. Table 3-4 is the average results of Γ at score = 50% by assuming a one-peak concentration on one of the incorrect responses. A three-level quantization scheme is also proposed in the table.

Table 3-4. Quantization of Γ (score = 50%)

Concentration on one choice	Average Γ	Quantization Range	Level
40%	0.15	0-0.2	L
60%	0.35	0.2-0.4	M
75%	0.55	0.4-1.0	H

Although Γ has the advantage of being independent of the score and it also provides more direct information on the incorrect responses, the measure of the total concentration is still important especially when evaluating the overall model condition. The score represents the contribution from the correct model and it should be included in the model evaluation. Therefore in order to properly model the student responses, we often need to consider both ***C*** and Γ for different aspects of the data. The details of the transformation between Γ and ***C*** are discussed in Appendix B.

Analysis with FCI Data

To see how these methods work in practice, I analyzed some FCI data. The data is taken from introductory physics classes consisting of 14 UMd and 2 PGCC classes.[4] The students are mostly engineering majors. Seven of the UMd classes are tutorial based and the other nine are with traditional instruction.[5]

Initial State

First the pre-instruction FCI data of all 16 classes were analyzed with the three-level modeling schemes described in the previous sections. The results are very similar for all classes. This is expected since the background of the incoming students is similar. Therefore, the results of the pre-data analysis are combined.

Table 3-5 is a list of the coding of the pre-test response types for all 29 questions on the FCI test. To avoid bias generated by variations (e.g. size) of the individual classes, the results were obtained by putting together all the student data from different classes rather than averaging the results of individual classes. The response types are obtained with the quantization scheme in table 3-3. The S-***C*** values of all questions are listed in table 3-5a.

Table 3-5a. Score and concentration values of pre-instruction FCI response (UMd)

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
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
	F16	**F17**	**F18**	**F19**	**F20**	**F21**	**F22**	**F23**	**F24**	**F25**	**F26**	**F27**	**F28**	**F29**	
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	

As shown in table 3-5b, the student responses are 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 favor a common incorrect model. More interesting results come from LM and LH types, which are strong indications for the existence of common incorrect models.

Table 3-5b. Pre-instruction FCI response types (UMd)

Types	LL	LM	LH	ML	MM	MH	HH
Patterns	No peak	Two peaks	One peak	Weak one-peak	Two peaks	Weak one-peak	One peak
Questions	15, 24	5, 9, 18, 28	2, 13, 22	3, 7, 21, 26	6, 8, 11, 14, 17, 20, 23, 25	12, 16, 29	1, 4, 10, 19, 27

A look at the details of the questions suggests that most of the questions with LM and LH types deal with two physics concept domains, the relation of force and motion and Newton III (see chapter 2 for details on the two concept domains). To further study the student understanding on these two physics concepts, we used our knowledge of student models and the specific distracters to select questions, which correspond to these concepts.

Table 3-6 shows the incorrect choices that attract a large percentage of student responses. As we can see from the individual questions (see Appendix A), the incorrect responses on questions in the Force-Motion group represent the physical model, which is defined in chapter 2, that a force is always needed in the direction of motion. The incorrect responses on questions in the Newton III group represent the dominant agent model that object with greater mass (greater velocity, etc.) exerts larger forces in an interaction.[6] (also see chapter 2 for more detailed discussion on the related incorrect student models)

Table 3-6. Student responses on the two concept groups of FCI test (UMd)

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

From table 3-6, in the Force-Motion group, before instruction about 60% of the students favor the idea that a force is necessary to keep an object moving. Similar situations are found in the group of Newton III questions where student responses usually have low scores and medium to high concentrations (LH or LM type) on the major distracters corresponding to the incorrect student models in the concept group.

With low scores and also low concentration (LL type), questions 15 and 24 reveal another kind of situation where the students are not particularly concentrated on any single response. Interestingly, both of the questions require detailed descriptions of physical processes that require an integration of various pieces of physics knowledge.

The above analysis shows that when students have certain common incorrect models on the physics concepts being tested, their responses on the related multiple-choice questions are often in LH or LM type. When students have no dominant models, their responses often have low concentrations. Therefore, we can use the concentration factor as a tool to evaluate student modeling conditions or to screen the test results to identify questions that are likely to trigger common incorrect student models.

The S-*C* Plot Analysis

To present the data in a visual format, we can use the S- ***C*** plot. The initial states, final states, and the shifts can be represented with points and vectors on the S-***C*** plot. Each point in the graph represents the averaged result on one question from all students. Since the tutorial and traditional classes have very different shift vectors, the results from the two different groups of classes will be presented separately. It is also interesting to compare the differences. Figure 3-4a and 3-4b are the S-***C*** plots of all the data (pre and post) from the tutorial and traditional classes. Each data point represents the result of all the student responses on one question. The vectors represent the shifts of the averaged pre and post results of all 29 questions for all students.

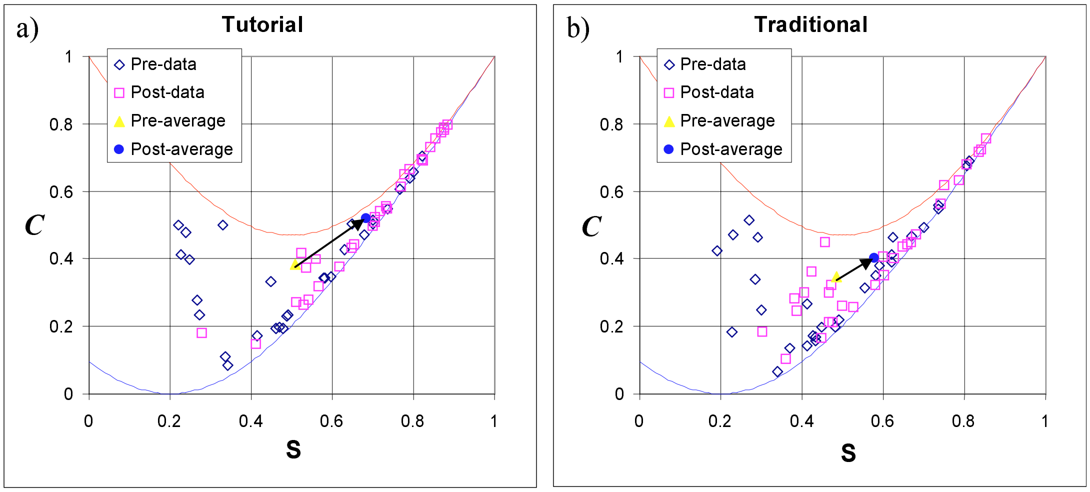


Figure 3-4. S-*C* plot of the overall performance of the tutorial and traditional classes

It is easy to see that the pre-states for both classes are very 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 table 3-4b, the 29 questions can also be separated into three groups based on student pre-instruction scores – high, medium, and low.[7] Since the high group is very close to the favorable situation, the low performance group should have a much larger contribution on the overall improvement. Therefore the shift of the low performance group should reveal 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 3-5, the S-*C* relation of these nine questions is plotted. 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 a incorrect model.

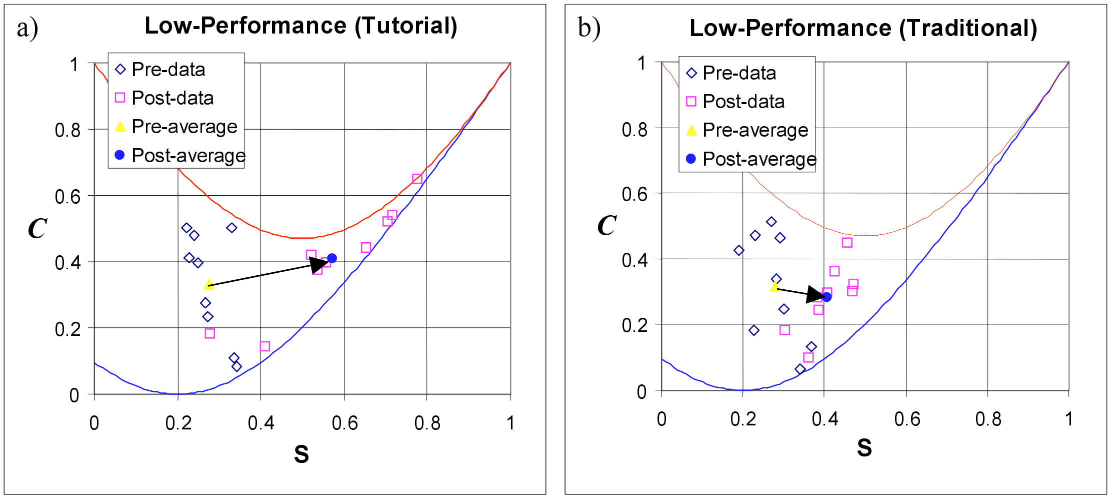


Figure 3-5. S-*C* plot for the low-performance group of questions

We can also study the details of student behavior in different concept groups. In figure 3-6, the shift of the questions in Force-Motion group is plotted.

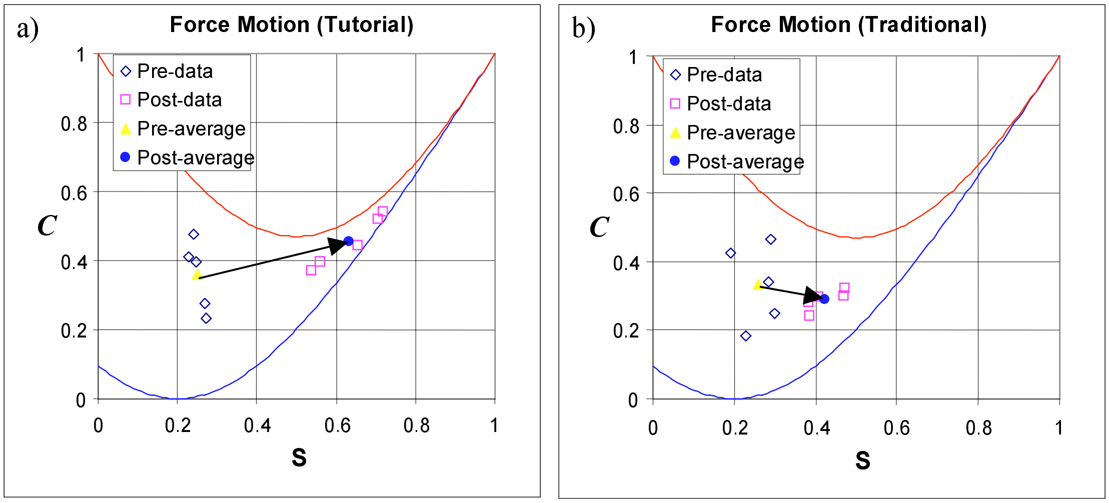


Figure 3-6. S-***C*** plot of the Force Motion group

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).

On an S-***C*** plot, the scatter of the data points for responses on questions in a concept group can be one evaluation on the consistency of the student performance. From figure 3-6, it is easy to see that the scores for the initial responses on the five questions are similar but the concentrations are very different. The scattering of the data points is mainly caused by the different concentrations. Therefore to evaluate the consistency of the student responses in a question group, we need to look at both the scores and the concentrations.

When working with the data, there are a few things that need attention. To obtain a result on one question for multiple classes, we have to group all the students together to calculate the score and concentration. Since different classes can have different numbers of students with different backgrounds, averaging over classes can give misleading results. It can also create data points outside the allowed region. Similarly, the averaged result for different questions could also be out of the allowed region, especially when the data points for these questions are more widely scattered and close to the upper boundary. For example, if we have two data points, (0,1) and (1,1) (LH and HH), the average is at (0.5,1) which is outside the allowed region.

The S-Γ Analysis

The S-***C*** plot shows the student overall modeling situations. 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 3-7. An interesting result is that the Γ's on low performance question are consistently higher than that of mid and high performance questions independent of the types of instructions and if it is pre or post data. Since high Γ's indicate strong distracters, it can be inferred that the students are tending to pick the same distracters in the low performance FCI questions. This also implies that these distracters may reflect responses produced by common incorrect student models.

Table 3-7. The averaged values of Γ for different performance groups based on the averaged score of the pretest

	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 3-8. FCI questions grouped based on pre-Γ values

High ($\Gamma > 0.4$)	#	**2**	**5**	**9**	**11**	**13**	**16**	**18**	**19**	**22**	**28**	mean
	S	0.33	0.25	0.27	0.45	0.22	0.65	0.23	0.82	0.24	0.27	0.37
	Γ	0.92	0.64	0.40	0.61	0.77	0.86	0.64	0.75	0.76	0.47	0.68
Mid (0.2~0.4)	#	**3**	**6**	**10**	**12**	**14**	**17**	**23**	**27**	**29**		
	S	0.42	0.58	0.80	0.62	0.59	0.68	0.58	0.77	0.70		0.64
	Γ	0.20	0.24	0.35	0.38	0.38	0.23	0.27	0.34	0.28		0.30
Low ($\Gamma < 0.2$)	#	**1**	**4**	**7**	**8**	**15**	**20**	**21**	**24**	**25**	**26**	
	S	0.74	0.74	0.46	0.60	0.34	0.49	0.43	0.34	0.41	0.48	0.50
	Γ	0.08	0.09	0.14	0.13	0.15	0.12	0.13	0.04	0.11	0.07	0.11

To confirm these implications, we need to look at the details of the questions. In table 3-8, the questions are regrouped based on the pre-Γ values. From the calculations, it is easy to see that the questions with high Γ's have significantly lower scores except for questions 16 and 19. If we leave out questions 16 and 19, the remaining questions in the group are the same questions with LM and LH type in table 3-5. This result also agrees with the implication that the poor scores on these questions are likely to be caused by incorrect but popular student models.

Questions 16 and 19 represent a different situation where most students picked the correct choice before instruction (high scores) and the majority of the remaining students picked a same distracter. This indicates that students do not have a common incorrect model on issues related to this question.

When comparing the pre and post results, the S-Γ plot is a good tool to illustrate the shift of student behavior on incorrect responses. Figure 3-7 is the S-Γ plot for all 29 FCI questions with pre and post data.

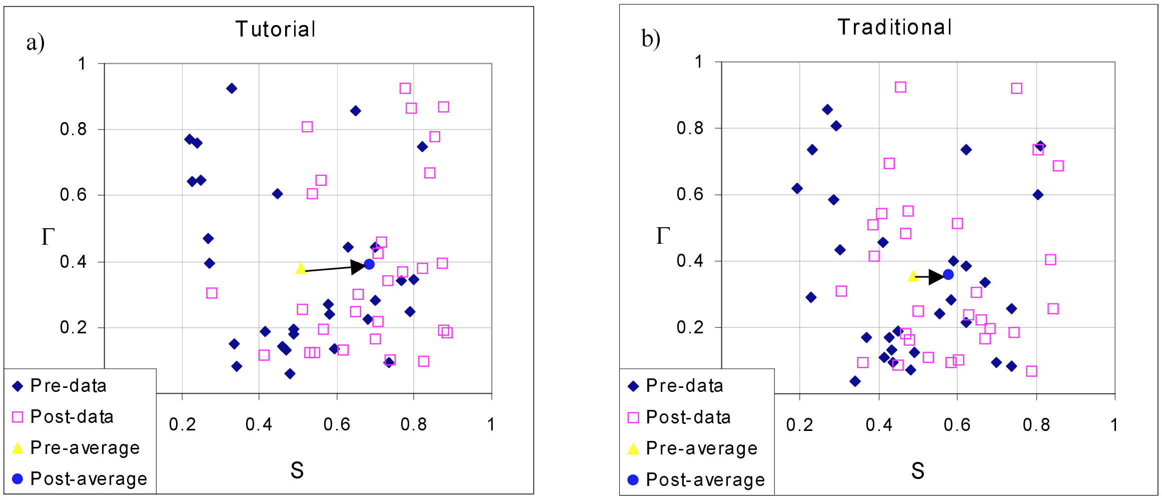


Figure 3-7. S-Γ plot for all 29 questions in FCI test

One advantage of the S-Γ plot is that Γ is not affected by score. From figure 3-4, the concentration of student post-instructional data gets much larger contribution from the scores and does not show much additional information. With similar scores, the concentrations on different questions often have similar values. On the other hand, from figure 3-7, the student post Γ's are still quite scattered just as the pre-instruction data. This implies that the students giving incorrect responses behave rather similarly before and after instruction (the group of students may not be the same ones). Therefore, to look for the information on students giving incorrect answers, we need to use the S-Γ plot.

In figure 3-8, the results of the low performance questions are plotted with the shift vectors of all the questions.

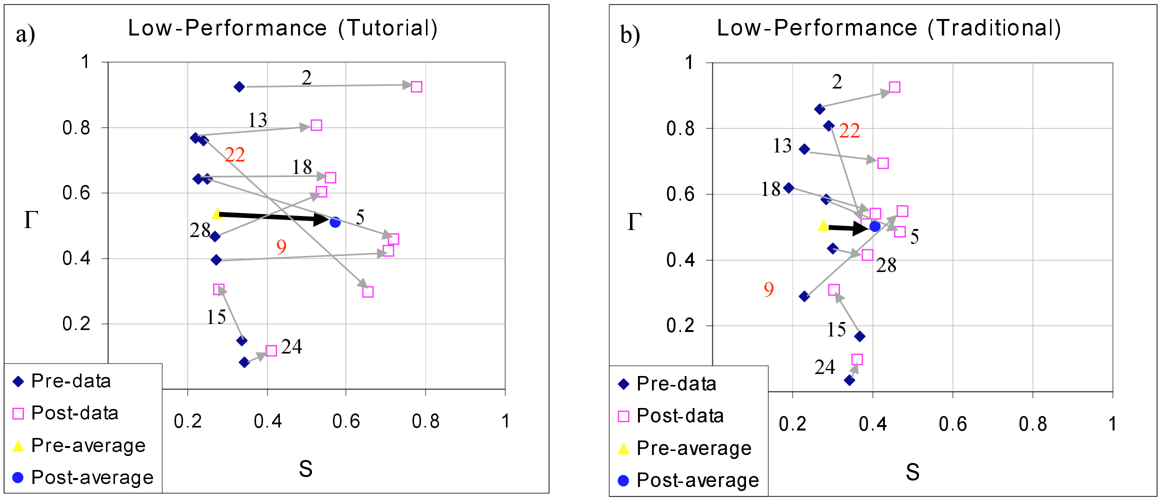


Figure 3-8. S-Γ plot for low performance group

The data in figure 3-7 and 3-8 also shows that the low performance groups have larger Γ's than the higher performance groups.

As demonstrated by the numerical results in table 3-7, the concentration of the incorrect responses is similar within the same question group (low, mid or high) for both types of instruction and pre or post. This implies that the students giving incorrect responses have the same kind of behavior on these questions before and after the instruction. From figure 3-8, we can see that not only the averaged results, but also the shifts of individual questions are similar except for questions 9 and 22. The detailed student responses on the two questions are listed in table 3-9.

Table 3-9. Student responses on questions 9 and 22 (the correct choice is shown in bold and the major distracter is italicized.)

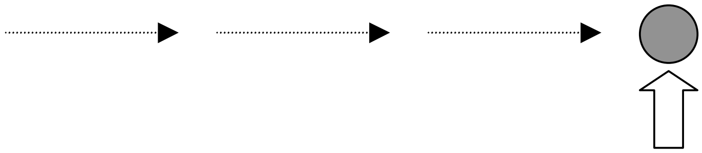
Question	9						22					
Choice	a	b	*c*	**d**	e	Γ	a	b	*c*	**d**	e	Γ
Tutorial												
Pre	0.07	0.17	*0.47*	**0.27**	0.01	**0.40**	0.03	0.06	*0.67*	**0.24**	0.01	**0.76**
Post	0.05	0.05	*0.2*	**0.70**	0.00	**0.42**	0.10	0.03	*0.2*	**0.66**	0.02	**0.30**
Traditional												
Pre	0.06	0.27	*0.42*	**0.24**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

The data shows that for question 9 (see figure 3-9 for details of the question, 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 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, see Appendix A for details of the question).

This result indicates that students after traditional instruction 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.

For question 22 (see figure 3-10), the data shows only one major distracter (question 9 has "b" and "c"). The variations of student responses on other distracters are around 5%. Therefore, Γ mostly depends on the student response on the main distracter. When students get significant improvement, as it is in the tutorial classes, the post-Γ will be significantly lower than the pre-Γ. Students with traditional instruction have much less improvement and the post-Γ is quite high.

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.

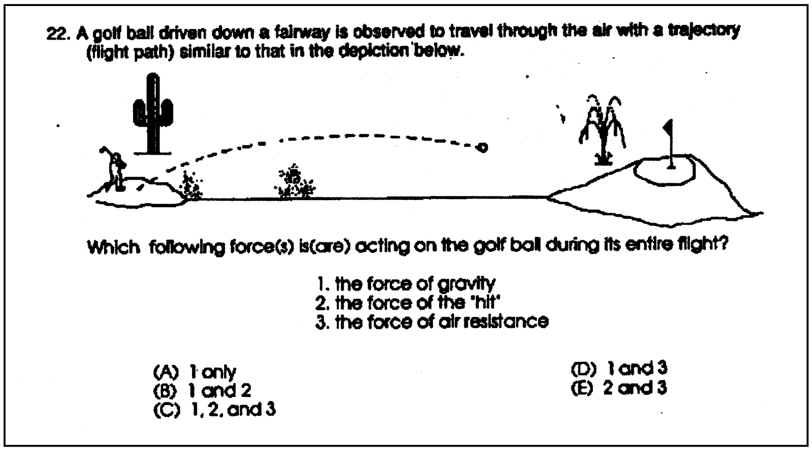
(B) the downward force of gravity and the horizontal force of momentum in the direction of motion.

(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 3-9. FCI question No. 9



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 3-10. FCI question No. 22

For 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 decrease on all the incorrect responses. In traditional 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 data remain similar (~ 15% changes).

Applications of Concentration Factor

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 well-designed multiple-choice test, we can use the concentration factor to evaluate student performance and their modeling conditions.

Facilitate Test Development

In PER and education research on other topics, more and more researchers are working on developing good multiple-choice tests.[8] To design a good test, the development should be based on systematic research on student understandings of the related topics. Once a prototype is proposed, it has to go through field tests and several round of revisions. In this process, we can use the concentration evaluation to help the development.

1. 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 that correspond to these student models. Using the concentration factor to analyze the results of the test, we can obtain quantitative evaluations and evidence on whether these distracters correspond well with the student models, and/or if the student models detected in interviews reflect do reflect common models.[9] If a distracter is effective, it will produce a low score but high $\boldsymbol{C}$ and Γ.

2. Detect items where a relevant distracter may be missing.

When the result shows a type of low score and low concentration, it indicates that the distracters are not attractive. This can be caused by three possible situations:

- None of the distracters reflects a common student model.
- For the context of the question, there does not exist a common student model.
- All the choices correspond well with the student models, and the number of the students in class with the individual models are almost an even distribution over all the models.

As a result, when a question with low concentration is detected, we need to do more research on the concept related to the question to further clarify the details involved.

Facilitate Instruction and Evaluation

In instruction, when we have a research-based test available, we can use the concentration factor to do more comprehensive evaluation on student performance and on the effectiveness of instruction.

Traditionally, student performance is evaluated with scores on a test. The problem with it is that when students have low scores, which often occurs in a pre-test situation (also possible for post-test with ineffective instruction), the information on how the majority of students get a question wrong can not be reflected with scores. This information is often the most important clue for instructors to improve the teaching strategies.

With the concentration factor, we can retain at least part of the information on students giving incorrect answers and infer the student modeling conditions. Especially when instruction is integrated with research, we can use concentration factor to evaluate:

- student modeling on different concepts,
- and student improvement with different instructions.

To use the concentration factor, several tools have been developed including the numerical evaluation of $\boldsymbol{C}$ and Γ, the S-$\boldsymbol{C}$ plot, and S-Γ plot. It is important to look at the different measures when considering student performance especially if we want to study the modeling issues. Since it is a multi-dimensional problem, each measurement can only provide us a glimpse from a certain angle. In order to get a more complete picture of the problem, it is needed to look at it from all possible angles and look for the information that best fits our goals.

Implications for Test Design

To use this method effectively, we need to design appropriate questions. Since we want to 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 physical model.

The number of choices in each question is an important factor. A small number of choices on each question can generate large distortion on student responses. On a multiple-choice test, student responses are constrained by the existing choices. If there is only a small number of choices, the random noise created by student guessing is large. In addition, for students with their own reasoning, if the choices do not allow them to have the opportunity to display their different understanding, these students will also be forced to make a guess. Therefore, we need a good number of choices carefully designed to match the various student beliefs (which should be research based).

To minimize the random distortion, it is helpful to allow a substantial number of choices in a question. These act like a noise space that is free for the students to get into so that the student responses will not be forced into any choices that are associated with student models and therefore can reflect more of their true understanding of the corresponding physics concepts.

In addition, with small number of choices (<3), the multiple-choice question becomes more like a true-or-false question. It is then meaningless to use this concentration evaluation, since once the score is known, the student incorrect responses are also obvious. Therefore, from our empirical experience, it is suggested that the number of choices for each question should be no less than 5.

Furthermore, to keep consistency in calculating the concentration, it is recommended to make all the questions 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 only result in insignificant changes (see chapter 5 for application examples with FMCE).

Summary

In this chapter, I have discussed a newly developed method to study the structure of the student responses in a multiple-choice test, which provides more useful information on the distribution of the student responses. The results can be further used to analyze the conditions of student mental models. Sample applications with FCI data confirm many widely recognized results and the additional information obtained with this method provides us new ways to study the student difficulties. We believe that this new method can be a more comprehensive evaluation on student performance in multiple choice tests.

References and Endnotes:

[1] The definitions of physical model is discussed in chapter 2. If the reader has skipped that chapter, it is recommended to review the section, *The Dynamics of Student Models*, where the definition is given.

[2] See chapter 2 for details and references of this test.

[3] Since this is very close to the random situation where the effect of the random variation is large, it will be difficult to differentiate whether the individual response is due to systematic reasoning with many different models or guessing.

[4] Data collected by Dr. J. Saul at the University of Maryland (UMd) and the Prince George Community College (PGCC).

[5] L. C. McDermott, P. S. Shaffer, *Tutorials in Introductory Physics* (Prentice Hall, New York NY, 1998)

[6] I. A. Halloun and D. Hestenes, "Common sense concepts about motion", Am. J. Phys. **53**, 1056 (1985).

[7] Low performance group: 2, 5, 9, 13, 15, 18, 22, 24, 28
Mid performance group: 3, 6, 7, 8, 11, 12, 14, 16, 17, 20, 21, 23, 25, 26, 29
High performance group: 1, 4, 10, 19, 27

[8] To number a few: In Maryland, the PER group is currently developing tests on electric conductivity and quantum mechanics. B. Hufnagel, a collaborator of the PER group at UMd, is using this tool to revise the astronomy diagnostic test (ADT).

[9] It can be argued that one can do the same thing by counting the numbers of responses on individual choices. It is an absolutely valid method, however, I would prefer to use the concentration because the counting number method requires more "RAM" and processing time from the reader (suppose we have 50 questions each with 10 choices which gives a 500-element table for one to read). It is also difficult to make direct comparison on results of different questions either within a test or from different tests.