Chapter 5: Refinements and Extensions

Introduction

In the previous chapters, I introduced the basic algorithms in model analysis to model and evaluate student learning of physics. Detailed examples with FCI data have been discussed extensively in chapter 4. In this chapter, I introduce two additional applications with model analysis. The first example is for the study of student models of mechanical waves based on a multiple-choice multiple-response question on wave concepts.[1] The second example is based on data from the FMCE (Force Motion Concept Evaluation) test, which also deals with the same topics of physics (introductory classical mechanics) as the FCI does. As we will see in later this chapter, these two examples involve additional issues in the process of "extracting the information" than those with a simple multiple-choice single-response test such as FCI. In responding to these new situations, further refinements and extensions have been made to the model analysis algorithms. In this chapter, I will also discuss a further development on analyzing student models with physical features. These examples are also used as demonstrations on how model analysis can be applied in various contexts.

Application with Student Models of Mechanical Waves: an example with a multiple-choice multiple-response test

The Wave Test is designed by Michael Wittmann at UMd to probe student understanding on concepts of mechanical and sound waves. It includes many questions of a variety of types. Here I will only analyze a single multiple-choice multiple-response (MCMR) question to show that a single question of this type allows the extraction of multiple student models. The question is shown in figure 5-1.

As indicated by research, there are three physical models involved:[2]

Model 1: Community wave model: the speed of the pulse depends only on the properties of the media (density, tension, etc.). (Correct)

Model 2: Particle-pulse like model, where students treat the wave pulses as particles thinking that the speed of a traveling pulse is affected by the amplitude of the pulse, etc. (Incorrect)

Model 3: Other irrelevant models and ideas and incomplete answers. (null model)

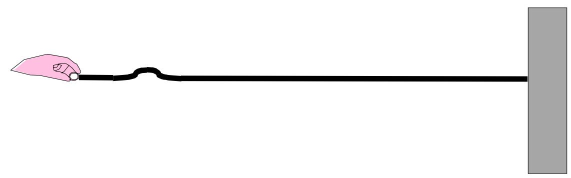
The data is collected from 4 UMd classes with similar instruction except that two classes had a tutorial on waves. The responses of the question can be modeled with table 5-1.

Table 5-1. Model scheme for the MCMR question in Wave Test

Models	Model 1	Model 2	Model 3
Responses	e, f, g, h	a, b, c, d, i, j	k

Wave Test Question

A long, taut string is attached to a distant wall (see figure). A demonstrator moves her hand up and down exactly once and creates a very small amplitude pulse which reaches the wall in a time t_0. How, if at all, can the demonstrator repeat the original experiment to produce a pulse that takes a less time to reach the wall.



Pick any correct statements from the following list.

a) Move her hand more quickly (but still only up and down once and still by the same amount).
b) Move her hand more slowly (but still only up and down once and still by the same amount).
c) Move her hand a larger distance but up and down in the same amount of time.
d) Move her hand a smaller distance but up and down in the same amount of time.
e) Use a heavier string of the same length, under the same tension
f) Use a lighter string of the same length, under the same tension
g) Use a string of the same density, but decrease the tension.
h) Use a string of the same density, but increase the tension.
i) Put more force into the wave.
j) Put less force into the wave.
k) None of the above.

Figure 5-1. The MCMR question in Wave Test developed at UMd

Since it is a multiple-choice multiple-response question, the formulation of the student model vectors is a little different. In the formulation, the whole question is still considered as one physical context. But with multiple responses, it is possible to use a single question to detect a mixed model state, although the resolution is low. In the calculation, the number of student responses corresponding to individual physical models cannot be used as model triggering probability in the construction of the single student model states. Because this question only forms one context instance and once a model state is triggered by this context, the student will use this model state to generate results. Therefore we can only get the information on whether the student has a pure model or is in a mixed model state. To obtain further details on the structure of the individual student model state requires larger number of questions.

The MCMR question is like a series of true/false questions constructed with one physics context. We can choose the weights on the corresponding physical models for a single student response vector to be either 0 or 1. The response vector for one student can be written as:

$$\mathbf{r}_k = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

where $y_\eta = 0$ when no response corresponds to the η^{th} physical model and $y_\eta = 1$ when at least one response is associated with the η^{th} physical model. The student model vector can be obtained with:

$$\mathbf{u}_k = \frac{1}{\sqrt{\sum_{\eta=1}^{3} y_\eta}} \begin{pmatrix} \sqrt{y_1} \\ \sqrt{y_2} \\ \sqrt{y_3} \end{pmatrix} \qquad (5\text{-}1)$$

Since the choice corresponding to the third physical model is a "none of the above" type, it is logically impossible to have any mixing between this model and the first two. Therefore, in this case there are only four possible student model vectors,

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Following similar procedures described in chapter 4, the class model states are analyzed and plotted in figure 5-2.

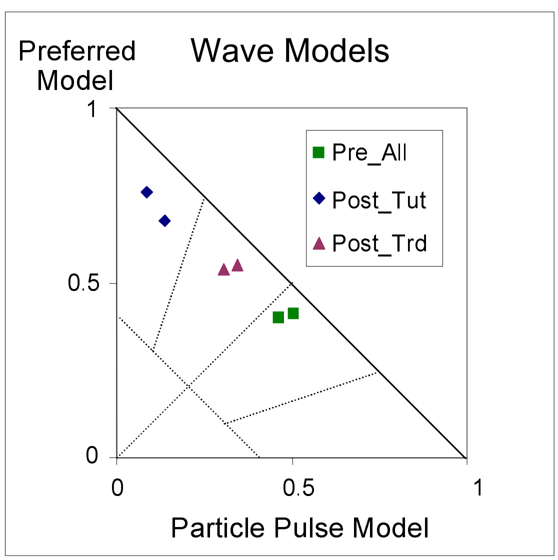


Figure 5-2. Student model state of waves

As we can see the initial situations are similar for all classes. The students all start with a mixed model state. For those students with the wave tutorial, the final states move into the favorable corner. The students without the wave tutorial remain in the mixed region. The numerical evaluation of student improvement for both types of is calculated in table 5-2.

Table 5-2. Student improvement on mental models of mechanical waves
($\mathcal{M}$ is the fraction of model improvement defined in chapter 4)

Classes	$\mathcal{M}$
Tutorial	0.62
Non-Tutorial	0.26

Application with FMCE Test

The FMCE test is developed by R.K. Thornton and D. Sokoloff at Tufts University.[3] It has 47 multiple choice questions on concepts related to diagrams (velocity, force, etc.), force motion, Newton III, and work and energy.[4] A copy of this test is included in Appendix A.

Concentration Analysis of FMCE Data

The student data used in our analysis is from California Polytechnic at San Luis Obispo (CalP).[5] The classes are calculus-based with two different instructions: traditional lecture with traditional lab, and *Real Time Physics* (RTP).[6]

- **Concentration Evaluation under Different Question Settings**

Unlike FCI questions where the number of choices for each question is almost the same (m = 5 except for question 16), the number of choices for FMCE questions varies between 6 and 9. The equations used to calculate the concentration factor of FMCE data will have different m's (see chapter 3 Eq. 3-2). We need to consider whether this will cause significant differences between analysis of questions having different m's. With this example, the concentration factor is used

- to evaluate the concentration features of all the FMCE questions and to compare them,
- and to compare the difference on the concentration between the FCI and FMCE questions.

To do so, we need to know what effect can be caused by this variation on different m's for different questions. We can get some insight if we use Eq. 3-2 in terms of the scaled length of a response vector, r_0:

$$C = \frac{\sqrt{m}}{\sqrt{m}-1} \times (r_0 - \frac{1}{\sqrt{m}}) = \frac{r_0 - \frac{1}{\sqrt{m}}}{1 - \frac{1}{\sqrt{m}}} \tag{5-2}$$

where

$$r_0 = \frac{\sqrt{\sum_{i=1}^{m} n_i^2}}{N}. \tag{5-3}$$

Notice that $\frac{1}{\sqrt{m}} \le r_0 \le 1$.

For the same r_0, with different m's, Eq. (5-2) will generate different ***C***'s. Figure 5−3, reveals the relation between the ***C*** and r_0 at different m's.

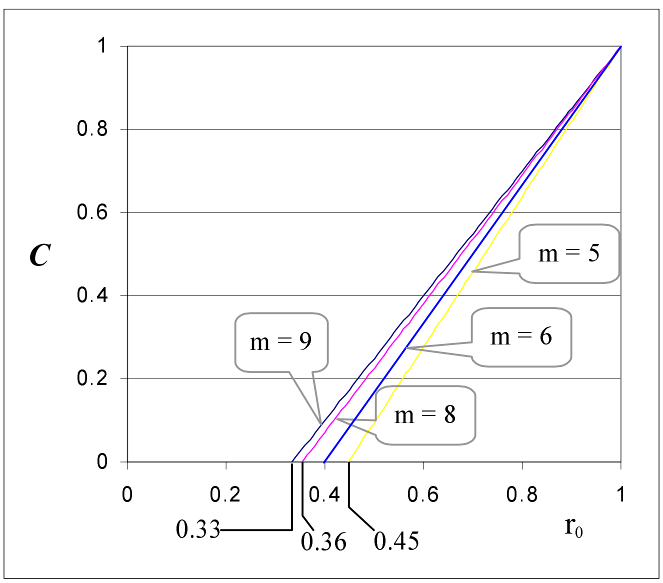


Figure 5-3. Comparison of concentration factors with m = 5 to 9

It is easy to see that for the same r_0 larger m's will produce larger ***C***'s. From Eq. 5-3, it is easy to see that the value of r_0 has major contributions from large elements of the response vector. For example, suppose we have a total of 100 responses on one question. Consider the following two response vectors shown in table 5-3.

Table 5-3. Examples for response vectors with similar distribution but with different m's

m	Response Vector (N=100)	r_0	***C***
9	$\mathbf{r}_1$ = (40, 30, 10, 7, 5, 3, 2, 2, 1)	0.519	0.278
5	$\mathbf{r}_2$ = (40, 30, 12, 10, 8)	0.530	0.150

As we can see, the distribution of the elements with large values for both vectors is similar and r_0 also gives similar results for the two vectors. For the case of $\mathbf{r}_1$, since the distribution occurs in a context with larger freedom on choices (larger m), the concentration factor in this case (m = 9) will be larger than the concentration of a similar distribution in a context with smaller freedom on choices (m = 5). Therefore, when evaluating a same response vector, the concentration factor takes into account the effect of all possible distributions for a particular question setting. On the other hand r_0 gives the evaluation on the distribution of the dominant elements of the response vector. As a result, when comparing the concentration of different response vectors with different dimensions (m's), we can consider using r_0 instead of $\boldsymbol{C}$. However, this alternative doesn't come without a price. When using r_0, the contribution from the dimensions with small elements is hardly reflected in the result. Thus, only when these dimensions are considered insignificant, i.e., hardly anyone will choose those choices, can we then use r_0 in our analysis.

Among FMCE questions, the different m's ($m \approx 8 \pm 1$) can create a maximum difference $\cong 0.1$ (between m = 6 and m = 9 at $r_0 \sim 0.4$) for the concentrations factors. For most questions, r_0 is around 0.6 and the error is less than 0.05, which is rather insignificant. Therefore, we can use Eq. 3-2 directly to calculate the concentrations of all FMCE questions and present them together.

With FCI questions, the number of choices in each question is significantly lower than that of the FMCE questions (m = 5 for FCI and m ~ 8 for FMCE). A direct comparison of the $\boldsymbol{C}$'s can create large errors. The alternative way is to use r_0. According to the data, student responses are mainly concentrated on three of the choices for most questions. Especially with FMCE questions, more than 90% of the responses are concentrated on about half of the choices. For each question, there are often 2~3 choices that almost never get chosen. Therefore, we will assume these choices insignificant and use r_0 to make comparison on the concentration of student responses on FCI and FMCE questions. An interesting effect is that when we ignore those insignificant choices in FMCE questions, the FCI and FMCE questions appear to have a similar "effective" dimension (number of effective distracters).

- **Concentration Analysis for FMCE Data**

The pre-instructional concentration factors for the FMCE questions are calculated and plotted in figure 5-4. From figure 5-4, it is easy to see that the 47 FMCE questions can be grouped into four groups (LH, LM, MM, and HH) based on the types of student responses (see chapter 3 for details on response types and quantization rules).

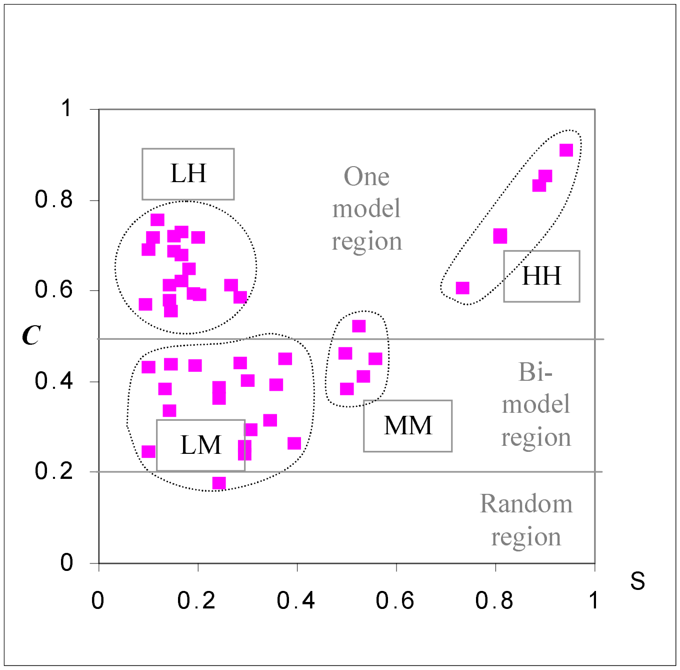


Figure 5-4. S-*C* plot of pre-instructional FMCE data
(each point represents the S-*C* result of a question)

As discussed in chapter 3, questions with LH type each has a strong incorrect attractor; questions with LM often have two incorrect attractors; questions with MM type have two attractors (one correct and one incorrect); questions with HH type each has a strong correct attractor. The individual questions in the four groups are listed in table 5-4.

Table 5-4. The concentration response types of the FMCE questions

Types	Questions	<S>	<*C*>
LH	1,2,4,8,9,10,11,12,13,14,16,20,28,30,34,36,38,44	0.17	0.65
LM	3,5,6,7,17,18,19,21,23,25,27,29,31,32,39,45,46,47	0.25	0.35
MM	22,24,26,35,37	0.52	0.45
HH	15,33,40,41,42,43	0.85	0.77

As we can see, 36 out of the 47 FMCE questions have LH or LM types of initial responses. This indicates that many students have strong incorrect initial models on the concepts related to these questions.

A simple comparison between the FMCE results shown in figure 5-4 and the FCI results in chapter 3 (see figure 3-5) reveals that FMCE questions seem to have higher concentrations than FCI questions. Especially for the low performance questions (low scores with LL, LM, and LH types), FMCE questions have lower average scores and higher average concentrations. See table 5-5.

Table 5-5. FCI and FMCE comparison on average results (pre-instruction) of low performance questions (score < 40%). The FCI results are from UMd and PGCC students and the FMCE results are from CalP students.

	FCI	FMCE
S	0.27	0.21
C	0.34	0.50
r_0	0.63	0.67

Since FCI and FMCE questions have significantly different m's, the results of ***C*** are in general not directly comparable. As discussed in the previous section, in this case, we can use r_0 as a standard to compare the results from FCI and FMCE tests. In table 5-5, the average values of r_0 for the low performance questions are also calculated. As we can see, the difference is small. This result indicates that for the low performance questions, although FMCE questions have a large number of choices, the distribution of the principal elements of student responses is similar to that from the FCI questions. This implies similarly attractive distracters in these low performance questions of both tests.

The distribution of the concentration for questions in both tests also reveals a significant difference between the FCI and FMCE questions. For FCI questions, only 27% of the questions have the LH or LM type of responses (pre-instruction). On the other hand, the pre-instructional data shows that 77% of the FMCE questions have LH or LM types of responses. Since questions with LH or LM type of responses are often related to the presence of strong incorrect student models, it appears that the FMCE test has more effective questions to probe student models.[7] To study the details of student models, we need to look at the structures of specific questions. In the following sections, I will do model evaluation with FMCE data and discuss more details on student models and the two instruments.

Model Evaluation of FMCE Test

Similar to the analysis of FCI questions in chapter 4, in this example I focus on two concept groups – the Force-Motion and Newton III. For the Force-Motion concept domain, there are many questions related to this topic in the FMCE test (questions 2,5,8-13, 14, 17, 21, etc). Most of these questions have LH type of responses (see table-5-4). The questions we choose to do model analysis on include questions 2, 5, 11, and 12. These questions are selected for two reasons: First, the questions all have simple responses without additional issues such as the interpretation of graphs. (Questions 14, 17, 21 also deal with force motion but the involvement of graphs may incur additional difficulties to students, i.e., students may be unable to express their models correctly with graphs.) Second, we try to use questions with different contexts (story-lines). As discussed in chapter 2, the model triggering depends heavily on the context of the questions. Diverse contexts/story-lines allow students to fully manifest their models. In the FMCE test, questions 8-13 all deal with similar physical contexts. Using only questions from this group is not a favorable setting for detecting mixed student model states. Therefore, questions 2, 5, 11, and 12 are selected to make the contexts as diverse as possible.

For the Force-Motion concept, the physical models are all based on the "Force-Velocity" relation, i.e., only one physical feature, the velocity, is involved. The questions also have either a one-to-one or a many-to-one correspondence between the choices and the physical models. Therefore, we can use the item-based modeling method to model these questions. The corresponding models and the responses to the questions are listed below and in table 5-6.

Model 1: It is not necessary to have a force to maintain motion and there is no such thing as a "force in the direction of motion". (Correct)

Model 2: A force is needed to maintain motion. This model also includes the ideas that there is always a force in the direction of motion and that the force is directly related to the velocity of motion. (Incorrect)

Model 3: Other ideas and incomplete answers. (Null model)

Table 5-6. Model scheme for Force-Motion questions in FMCE test

Question	Model 1	Model 2	Model 3
2	D	B	others
5	D	B	others
11	A	G	others
12	A	D	others

Following similar procedures as in chapter 4, the student class model states are calculated and plotted in figure 5-5. As we can see, this result is similar to the results from FCI data (see figure 4-11).

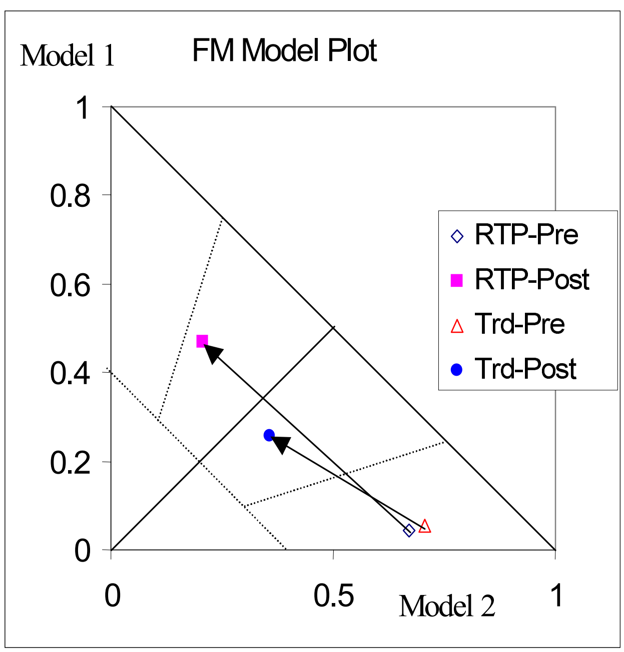


Figure 5-5. Student model plot on Force-Motion (FMCE)

Now we have seen the analysis of the same physics concept with two different tests. Since the questions are different, it is difficult to directly compare the scores from the two tests. One of the advantages of using model analysis is that even though the questions are very different, if they are designed to probe the same concept, the data can be transferred into the same model space. Then the obtained model states, which reflect both the student understandings and the features of the instruments, can be made useful to evaluate the different properties of both issues under different settings. For example, with a group of students with known background, we can study the model triggering features of different instruments; with a calibrated instrument, we can study the modeling of students with unknown background.

- **Combined Pattern Analysis for Modeling Newton III Questions**

There are 10 FMCE questions on Newton III (questions 30 – 39). Questions 35 – 39 are not used since two of them (36, 39) are identical to the related FCI questions (questions 11, 13), and we are interested in comparing questions with different structures. However, the two identical questions provide an opportunity to evaluate the two student populations with a similar standard. (This will be discussed later in this section.)

As discussed in chapter 2, the student models on Newton III involve multiple physical features – velocity, acceleration, mass, etc. Students can construct models with different emphasis on these physical features under different contexts. The combination of a variety of physical features in one question can make the situation much more complicated. For the FMCE questions used in our analysis (questions 30 – 34), two physical features are consistently involved all five questions:

1. Force – Velocity: Many students have a model that an object with a larger velocity will exert a larger force on an object with smaller velocity.

2. Force – Mass: Many students have a model that an object with a larger mass will exert a larger force on an object with a smaller mass.

Considering all possible combinations of both physical features, students can construct three incorrect physical models and one correct expert model. Therefore, for this group of FMCE questions, we can set up five different physical models as:*

M_0: Null model.

M_1: During the interaction, both objects always exert the same amount of force to the other regardless of either mass or velocity of the object. (Correct)

M_2: During the interaction, the two objects can exert unequal amount of force to the other and the one exerts the larger force doesn't depend on mass but it does depend on the velocity of the object. (Incorrect)

* The numbering system of physical models is different from what is used in chapter 2 and 4. This makes it convenient to study multi-dimensional mode space including sub-spaces (with physical features).

M_3: During the interaction, the two objects can exert unequal amount of force to the other and the one exerts the larger force depends on the mass but it does not depend on the velocity of the object. (Incorrect)

M_4: During the interaction, the two objects can exert unequal amount of force to the other and the one exerts the larger force depends on both the mass and the velocity of the object. (Incorrect)

The choices in FMCE questions often have one-to-many choice-to-model correspondence (one choice to many models). For example, question 30 describes a head on collision between a big truck and a car where both vehicles are moving at a same speed (see Appendix A or figure 2-8). The question asks the students to compare the force between the truck and the car during the collision. As we can see, students with the expert model (M_1) and students with the incorrect M_2 will all come up with the same response that the force has same magnitude, which is also the correct answer. Therefore, from the student response on this question alone, we can not determine which type of model the student might have. As a result, student responses on this group of FMCE questions cannot be coded with item-based modeling.

To deal with this situation, I introduce an additional modeling algorithm – *Combined Pattern Analysis* (CPA). It is developed based on the idea of looking for coherence in the student responses over different questions to extract the underlying models.

Consider if we have a series of questions all dealing with a same physics concept. Then different models will often generate different patterns of responses over a series of questions (assuming the questions are so designed.). For example, suppose we have five questions in a concept group with two models denoted as M_1 and M_2. Suppose a student with a consistent model, M_1, will create a response pattern of "ABBCA" and a student with a consistent model, M_2, will create a pattern of "ACCCB". Then it is obvious that question 1 and 4 (giving same answers for both models) cannot be used to determine the actual student models. But when we combine all the responses together and analyze the combined patterns, the student models can be measured based on the agreement of the student response pattern and the response pattern of a specific model. For the two-model example, this method doesn't show much advantage, since we can discard those questions with identical responses (if we have enough questions). But for a situation with more than two models, the overlap of responses can be in different places for different models. Then we may not be able to discard any questions and we have to match the entire pattern for different models.

- **The formulation of the combined pattern analysis (CPA)**

Suppose we have "m" multiple choice questions associated with *w* models. Each question contains "h" choices. Denote the different models as M_η, where $\eta = 0, 1, \dots$ w-1. For convenience, M_0 is used to represent the null model. So the total number of physical models equals w, which is also the dimension of the model space. Define $\mathbf{P}_\eta$ as the response pattern of a pure consistent model M_η. Then P_η can be written as

$$\mathbf{P}_\eta = (C_{\eta 1}, C_{\eta 2}, ..., C_{\eta m})^T \tag{5-4}$$

where $C_{\eta j}$ is the symbol representing the choice of the j^{th} question corresponding to the physical model M_η.

For the example here, the total number of possible patterns equals h^m. In general cases, the number of the patterns for the non-null models is w-1 (each model corresponds to a single pattern). Then the number of possible patterns associated only with the null model will be $h^m - w+1$. Notice that there is a probability in the order of $1/h^m$ that any given model pattern will appear even when the student is using a null model. This is only significant when "h" is small (≤ 3)). Since we won't study the details of the different possible configurations of the null models, we can still represent them with M_0. Therefore, M_0 will often be associated with quite a number of random patterns.

Next, let us define the student response pattern as $\mathbf{S}_k$, where k = 1, …, N represents the different students. Write $\mathbf{S}_k$ in the form of a vector

$$\mathbf{S}_k = (s_1^k, s_2^k, ..., s_m^k)^T$$

where s_j^k represents the answer of the k^{th} student on the j^{th} question.

Now I will introduce a logical operation "∘" that measures the agreement of two symbols. It is defined as

$$X \circ Y = \begin{cases} 1 & X = Y \\ 0 & X \neq Y \end{cases} \tag{5-5}$$

where X and Y are two arbitrary symbols. A vector operation can also be defined by

$$\mathbf{A} \circ \mathbf{B} = \sum (A_i \circ B_i)$$

This gives a value between 0 and the dimension of the vector. It can be used to evaluate the symbolic agreement of two vectors.

- **Modeling Student Responses with CPA**

Given a student response pattern $\mathbf{S}_k$, we can create the student model vector with the following procedures. First, $\mathbf{S}_k$ is compared to all the $\mathbf{P}_\eta$'s. If a perfect match is found, the corresponding M_η will be used as the student model. If no perfect match is found, then the agreement with each $\mathbf{P}_\eta$ will be stored in a weight vector defined as

$$\mathbf{G}_k = (G_0^k, G_1^k, G_2^k, ..., G_{w-1}^k)^T \tag{5-6}$$

where

$$G_\eta^k = \mathbf{S}_k \circ \mathbf{P}_\eta \qquad \eta = 1, \ldots, w\text{-}1$$

The G_0^k represents the contribution from the null model and it is equal to the total number of those elements in $\mathbf{S}_k$ that have not made contribution to any G_η^k with $\eta > 0$.

In some cases, some of the multiple-choice responses may be interpreted as matching more than one model, therefore, it is possible to have

$$\sum_{\eta=0}^{w} G_\eta^k > L$$

Therefore, we need to normalize the weight vector. Denoted the normalized weight vector by $\mathbf{g}_k$, where

$$|\mathbf{g_k}| = L$$

Since the overlapping (one-to-many) happens in the modeled part of the weight vector (G_η^k for $\eta > 0$), we need to normalize the modeled part of the G(k) separately:

$$\mathbf{g}_k = (G_0^k, \overline{G_1^k}, \ldots, \overline{G_{w-1}^k})^T \qquad (5\text{-}7)$$

and $\overline{G_\eta^k} = \dfrac{G_\eta^k}{\sum_{\eta=1}^{w} G_\eta^k}(L - G_0^k)$

Then the single student model vector, $\mathbf{u}_k$, can be obtained with

$$|\mathbf{u}_k\rangle = \frac{1}{\sqrt{L}}(\sqrt{G_0^k}, \sqrt{\overline{G_1^k}}, \ldots, \sqrt{\overline{G_{w-1}^k}})^T \qquad (5\text{-}8)$$

The steps to get class model density matrix and student model states will be the same as in model analysis discussed in Chapter 4.

- **Analysis of FMCE Questions with Newton III**

Questions 30 to 34 are used in our analysis. In table 5-7, the four physical models with the configurations of the corresponding physical features are listed. A binary coding is used to represent the different configurations where a value "1" ("0") stands for the existence of a dependent (independent) relation. The response patterns for the five questions are also listed.

Table 5-7. Student models and response patterns on Newton III question in FMCE test

	Mass	Velocity	Q30–Q34
M_1	0	0	EEEEE
M_2	0	1	EBBEB
M_3	1	0	AAAEE
M_4	1	1	AFBEB, AFFEB

As we can see, on question 33 students with all the different models will produce the same response. Therefore, this question is not included in later calculation. Using the CPA method, the pre-instructional student class model density matrix is calculated with data from questions 30, 31, 32, and 34. The density matrix is shown in table 5-8.

Table 5-8. The class model density matrix for FMCE questions on Newton III (pre-instruction)

	M_0	M_1	M_2	M_3	M_4
M_0	*0.075*	*0*	*0.031*	*0.045*	*0.061*
M_1	*0*	**0.086**	**0.042**	**0.044**	**0.065**
M_2	*0.031*	**0.042**	0.141	0.091	0.14
M_3	*0.045*	**0.044**	0.091	0.096	0.137
M_4	*0.061*	**0.065**	0.14	0.137	0.602

The density matrix has 5 dimensions with M_0 as the null model. From the density matrix, we can see that before instruction, the majority of the students are concentrated on the incorrect models ($M_2 - M_4$) and M_4 has the dominant population. It can also be observed that the off-diagonal elements involving M_1 are significantly lower than those involving M_2, M_3, and M_4. This indicates that students giving responses corresponding to the correct model (M_1) are quite consistent on using their model, which also implies that students using one of the incorrect models have a stronger tendency to switch between different incorrect models, i.e., the mixing within the group of incorrect models are much higher than the mixing between the correct model and the incorrect models.

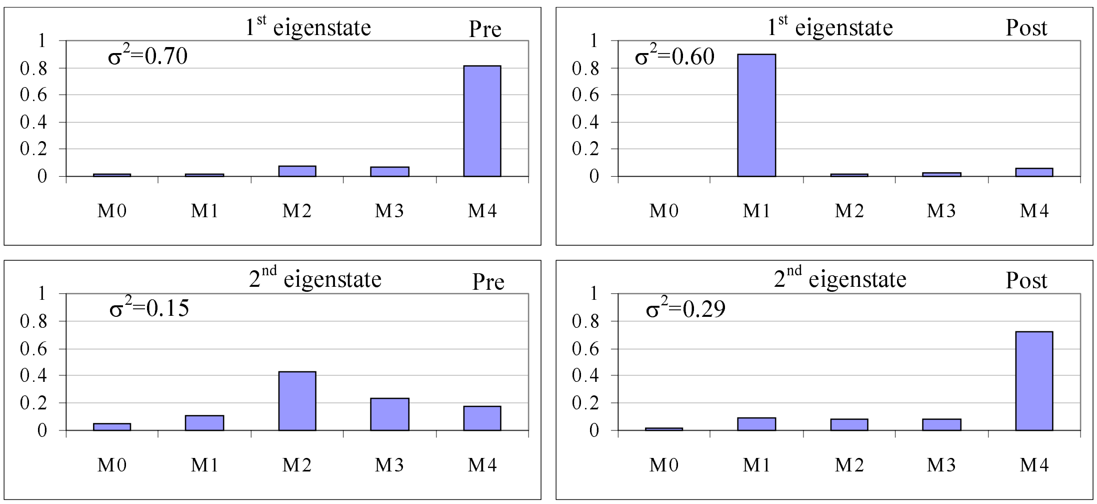
The class model density matrix for post-instructional data is also calculated and shown in table 5-9. The data is from classes with Real Time Physics. Data from classes with traditional instruction show little difference between pre and post test results and is not used here. All data are from CalP for the semester of spring 1999.

The results show that many students switched to the correct model (M_1) after instruction. The number of students using the incorrect models is significantly reduced. Again we can see that students with a correct model are quite consistent in using it and students with incorrect models are likely to switch between the different incorrect ones.

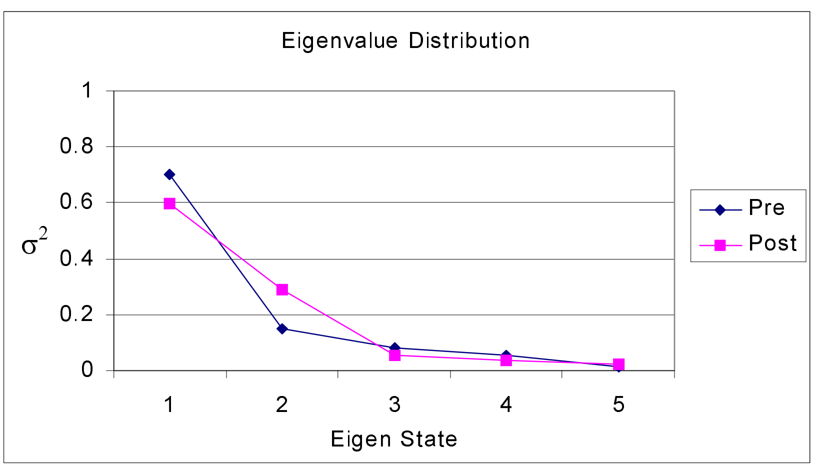
Table 5-9. The class model density matrix for FMCE questions on Newton III (post-instruction with Real Time Physics)

	M_0	M_1	M_2	M_3	M_4
M_0	*0.038*	*0*	*0.011*	*0.015*	*0.026*
M_1	*0*	**0.567**	**0.044**	**0.056**	**0.065**
M_2	*0.011*	**0.044**	0.07	0.044	0.074
M_3	*0.015*	**0.056**	0.044	0.067	0.08
M_4	*0.026*	**0.065**	0.074	0.08	0.258

To see more details on student class model states, the eigenvectors of the pre and post-instructional model density matrices are calculated and plotted in figure 5-6.



(a) Model elements of the first two eigenstates with largest and second largest eigenvalues



	Pre	Post
σ_1^2	0.70	0.60
σ_2^2	0.15	0.29
σ_3^2	0.08	0.05
σ_4^2	0.05	0.04
σ_5^2	0.02	0.02

(b) Distribution of eigenvalues for all five eigenstates

Figure 5-6. Model elements and eigenvalue distributions of the first two eigenvectors

The eigenstates are numbered according to the value of their eigenvalues (from large to small). Since the sum of the eigenvalues of the first two class model states is larger than 0.85, only these two model states are plotted in figure 5-6a (in model energy, see chapter 4). In figure 5-6b, the distribution of the eigenvalues for all eigenvectors are also plotted for both pre and post results. Here, we can also see that the students often have quite different model state (either correct or incorrect). Therefore, even though the eigenvalue for the primary state is comparatively small, we can still use eigenvalue decomposition to analyze the class (see chapter 4 for more details).

A comparison of the primary model states shows clear concentration on M_4 for pre data and M_1 for post data. The secondary model for pre-instructional data shows more mixing in between the incorrect models. The post-instructional models indicate very consistent M_1 and M_4 although the mixing among the incorrect models is still higher than that between the correct and incorrect models.

As implied by the analysis, M_2 and M_3 are not very popular (small diagonal elements) and students with these two models are more likely to mix with M_4 than students with M_1. Therefore, we can collapse M_2, M_3 and M_4 into one incorrect model denoted as M_4' and form a 3–D model space with M_1(correct), M_4'(incorrect) and M_0 (null model). As we can see, this result agrees with the dominant agent model discussed in chapter 2. The 3-D model space can significantly simplify the calculations and makes it easier to compare with the results from other 3 model examples such as FCI results. Since question 33 does not distinguish the models, it is removed from our data set.

The construction of the combined 3-D model space is shown in table 5-10. As we can see, M_4' (the major incorrect model) includes the elements from M_4, M_2 and M_3. The details of the transitions between these incorrect models are fine structures within the general M_4' state and will not be discussed in this dissertation. Here the calculation is focused on student model mixing between M_1 and M_4'.

Table 5-10. Reorganize the 5-D model space (N3) of the FMCE questions to a 3-D model space

	Vector	Notes	Q30, 31, 32, 34
M_4'	100	Incorrect	AFBB, AFFB EBBB, AAAE
M_1	010	Correct	EEEE
M_0	001	Noise	Other

The three new models can then be described as:

M_0: Null model
M_1: During the interaction, the force does not depend on either mass or velocity of the object. (Correct)
M_4': During the interaction, the force is dependent on mass, velocity or both. (Incorrect)

Based on the patterns in table 5-10, we can use CPA again to model the student responses. Notice that there are 4 patterns corresponding to M_4'. In the calculation, first, student responses are matched with all four patterns. If a perfect match is found, the student will be assigned with M_4'. If no perfect match is found, the largest weighing factor from the four matches is used as the weight for M_4'. Define the 3-D student model response vector as $\mathbf{r}_k$, we can write

$$\mathbf{r}_k = (r_0^k, r_1^k, r_2^k)$$

$$r_0^k = G_0; \quad r_1^k = \frac{4 - G_0}{Max(G_2, G_3, G_4) + G_1} \cdot G_1;$$

$$r_2^k = \frac{4 - G_0}{Max(G_2, G_3, G_4) + G_1} \cdot Max(G_2, G_3, G_4)$$

where the function Max() returns the maximum value from the input variables. The corresponding student model vector $\mathbf{u}_k$ is

$$\mathbf{u}_k = \frac{1}{\sqrt{m}}(\sqrt{r_0^k}, \sqrt{r_1^k}, \sqrt{r_2^k})$$

With the three-model space, student data can be analyzed similarly as described in chapter 4. The student model states are plotted in figure 5-7.

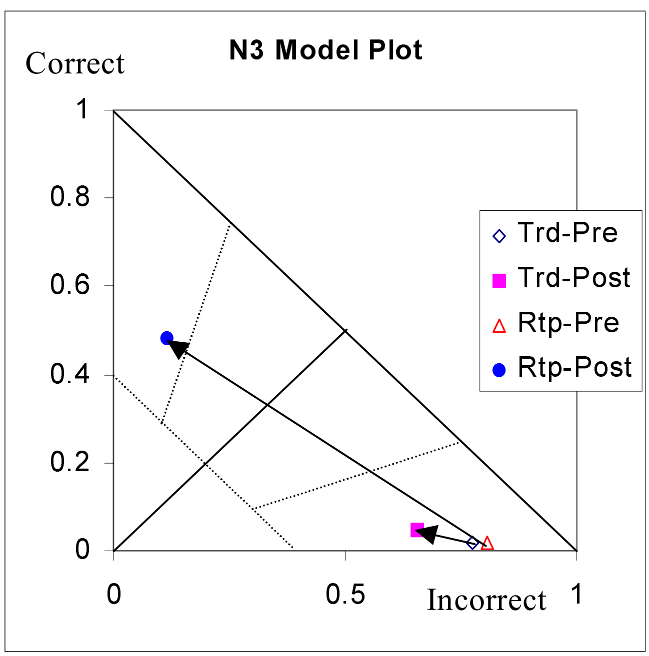


Figure 5-7. CalP student model plot on Newton III calculated with FMCE questions 30, 31, 32, and 34

As we can see, both groups of students start at a very similar situation. The students with Real Time Physics labs have a greater improvement toward the correct physical model than the students with traditional labs. The fraction of the possible model gain attained is also calculated in table 5-11.

Table 5-11. The fraction of possible model gain on Newton III (FMCE)

	$\mathcal{M}$
Traditional	0.077
RTP	0.627

From the results, the model-based performance of the students with RTP is similar to that of the students with tutorials found with FCI data in chapter 4. These classes also have similar possible model gains. However the absolute final model states are still much less favorable than the results from FCI. For students with traditional instructions, the performance is much poorer on the FMCE.

Since these results are from two different groups of students under different instructions, and the data is obtained with different tests, there are many uncertain factors that may cause the difference. As pointed out earlier, there are two identical questions in both tests. Therefore, we can use these two questions to find out the difference between the two groups of students.

- **Comparison of Student Performance on FCI and FMCE**

Using the two identical questions (FCI 11/13 and FMCE 36/39), the class model states for the two groups of students (UMd and CalP) are calculated. The same 3-D model space discussed in chapter 4 is used and the student responses are coded with item-based model assignment (see table 5-12).

Table 5-12. Modeling scheme of the two identical questions on FCI and FMCE test

	Question	Correct	Incorrect	Null
FMCE	36	a	b, c	d, e, j
	39	e	b, d	a, c, j
FCI	11	e	b, d	a, c
	13	a	b, c	d, e

The results of the class model states for the two groups of students are shown in figure 5-8. For easy comparison, the class model states obtained with FMCE questions 30-34 are also plotted (these states are the same ones in figure 5-7).

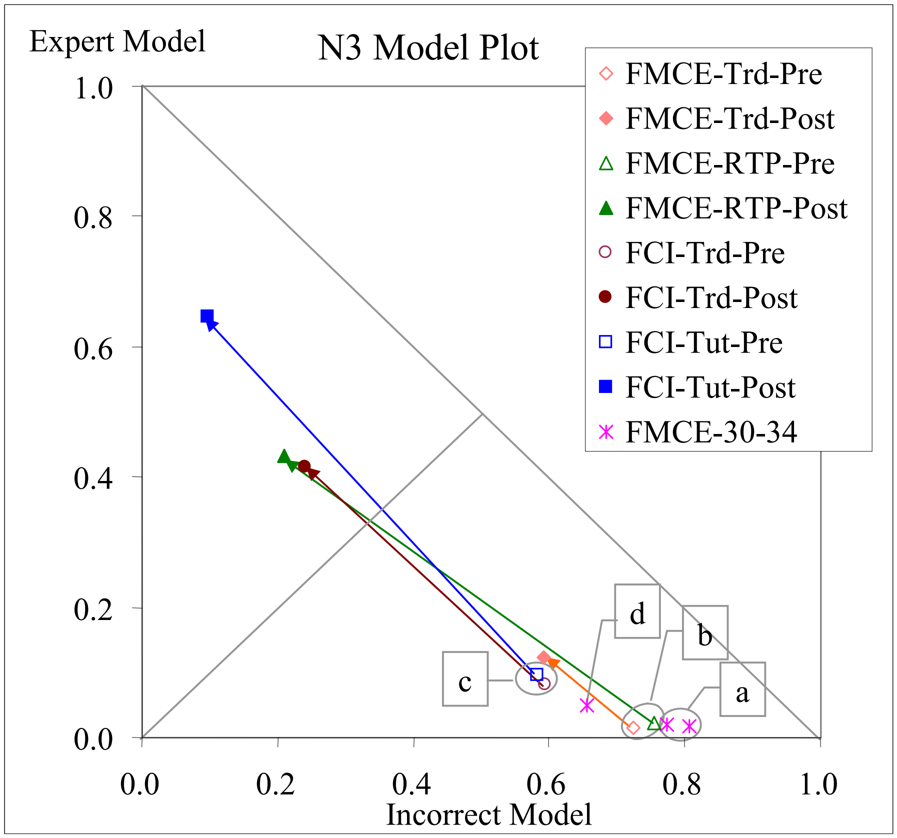


Figure 5-8. Comparison of class model states on N3 under different situations. All states with arrows showing the pre and post results shifts are obtained with the two identical questions. The FCI data is taken from UMd. The FMCE data is from CalP. The question groups are: A (FMCE questions 36 and 39), B (FMCE questions 30, 31, 32, and 34), C (FCI questions 11 and 13). Group A and C are identical.

a. The two initial states shown in figure 5-7, which is obtained from FMCE questions (group B).
b. The two initial states obtained with the two identical questions, FMCE questions (group A), from the same group of student as in “a”.
c. The two initial states obtained with the two identical questions, FCI questions (group C), from UMd students.
d. The post-instruction model state of CalP students with traditional instruction (obtained with FMCE questions – group C).

The results show that students from the two schools have noticeably different initial states (see group “b” and group “c”). This indicates that the backgrounds of the students from the two schools are different. Students at CalP have a more dominant incorrect initial state (larger eigenvalues which can be inferred from the shorter distance between the model states and the upper boundary line). In addition, the model states are also more consistent (less mixing). These results indicate that compared to UMd students, there are more CalP students having quite consistent incorrect initial models on N3.

With two different sets of questions (group A – question 36 and 39, group B – questions 30, 31, 32, and 34), the students from CalP show comparatively similar initial states (see figure 5-8 for cases "a" and "b"). This indicates that both sets of questions give fairly consistent measurement on student models of N3.

Still, we would like to understand what different aspects might be probed by using the two different tests. To do so, we need to find two similar student populations. From figure 5-8, we can see that the final model state of the students from CalP with traditional instruction is almost identical to the initial states of students from UMd. Therefore, we can use the results from these two groups of students to compare the different features of the two tests.

Evaluation of Multiple-Choice Diagnostic Instruments

Model Triggering Properties

As discussed in chapter 2, student models are highly context dependent. The structure of a test question plays an important role in the student model triggering process. Our goal here is to measure the student model states, which are often mixed states. Therefore, how different settings of test questions can affect the student model triggering process and the measurement of the corresponding student model states are of great importance to researchers for both correct interpretation of the results and the design of effective instruments. In the following sections, I will introduce three mathematical tools that can provide assessment on different model triggering features of test questions.

Model Hopping – A differential evaluation of model triggering features

From figure 5-8, we can see that for the same students (CalP, traditional instruction), the two groups of questions (group A – question 36 and 39, group B – questions 30, 31, 32, and 34) generate quite different class model states from post-instructional data. Since the students are the same, the difference has to be caused by different settings of the questions. The state from group B indicates a more consistent incorrect model. The details of the questions in group B (see Appendix A) reveal that all four questions in the group deal with a similar context/story-line of "car – truck collision" and the questions are all grouped together. The two questions in group A deal with two different contexts. In addition, the two groups of questions also deal with different physical features. Group A deals with the source of the force and the mass issue, while group B deals with the mass issue and the velocity issue. This can also contribute to the difference.

Since the models in N3 involve multiple physical features, with the available questions, it is difficult to isolate effects caused by the individual physical features and the structure of the different story-lines. To simplify the study of the context issue, I choose the Force – Motion concept domain, which involves only one physical feature: velocity.

To study the model triggering properties of the questions, I will introduce a new measurement, the *model hopping frequency*. It can be evaluated in two ways. One way is to study the questions and the other is to study the students. Assume we have a set of questions all dealing with a same physics concept. When studying the questions, f_h is defined as the average probability that a particular question within a series of questions will trigger students to use a different model than he/she did on other questions. When studying a single student, define f_s as the single student model hopping frequency which gives the average number of times that a single student will change models to solve a set of questions. For example, with a set of five questions, if a student use models in the order of (1, 2, 2, 1,1), there is one change between the first and the second question and one change between the third and fourth question. Scaled with the largest number of possible model changes (= 4), f_s is found to be 2/4 = 0.5. Therefore, f_s can be calculated with

$$f_s = \frac{\text{Number of model changes}}{\text{Number of possible model changes}} \qquad (5\text{-}9)$$

Notice that f_s is always obtained with a definite set of questions in a given order. Even if only the order of questions is changed, f_s needs to be measured and calculated again, because the ordering of questions is also part of the contextual information that might result in a triggering of different models, especially when questions are grouped next to each other (e.g. FMCE questions 8 ~ 13).

With the above example, suppose the five questions are not grouped together and have different context structures so each question initiates an independent model triggering process (e.g. FCI questions 5, 9, 18, 22, 28). Assume the student doesn't change. Then we expect that we can reorder the questions without causing much difference to the data. If we switch the third and fourth question, the model used by the same student would be (1, 2, 1, 2, 1). The corresponding f_s is found to be 1. As we can see, f_s has strong dependence on the ordering of the questions. This can cause problems when studying "stand alone" questions each with reasonably independent model triggering process. In such situations, the order of the questions should not impose significant effect on the results.

To obtain a more complete picture of the model triggering features with emphasis on questions, I would like to introduce a model-hopping matrix $\mathbf{F}_h$ as

$$F_h = \begin{bmatrix} 0 & f_h^{12} & f_h^{13} & \cdots & f_h^{1m} \\ f_h^{21} & 0 & f_h^{23} & \cdots & f_h^{2m} \\ f_h^{31} & f_h^{32} & 0 & \cdots & f_h^{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_h^{m1} & f_h^{m2} & f_h^{m3} & \cdots & 0 \end{bmatrix} \qquad (5\text{-}10)$$

The dimension of this matrix is the number of questions in the question set, "m", and the questions are indexed from 1 through m. The element f_h^{ij} describes the model-hopping frequency between question *i* and *j*. Denote the model used by the k^{th} student on question *i* as M_k^i. For a total of *N* students, f_h^{ij} is calculated with

$$f_h^{ij} = \frac{\sum_{k=1}^{N} M_k^i \otimes M_k^j}{N} \qquad (5\text{-}11)$$

and

$$M_k^i \otimes M_k^j = 1 - M_k^i \circ M_k^j = \begin{cases} 1 & M_k^i \neq M_k^j \\ 0 & M_k^i = M_k^j \end{cases}$$

Obviously, the diagonal elements are always zero. When considering question ordering, in general $f_h^{ij} \neq f_h^{ji}$ and to determine the two elements, we need to actually carry out two measurement with two sets of questions in the different orders.

When question ordering is considered insignificant, we can use upper half of the matrix to store data and the overall average model-hopping frequency, f_h, can be obtained with

$$f_h = \frac{\sum_{i=1}^{m} \sum_{j=i+1}^{m} f_h^{ij}}{C_m^2} \qquad (5\text{-}12)$$

In the FMCE test, 6 questions (2, 5, 8, 9, 11, 12) are used to study the FM model. Questions 2 and 5 are based on one story-line but with two different settings on the motion (pushing a sled on ice surface). Questions 8, 9, 11, 12 all deal with a very similar story-line and context settings (tossing a coin and pushing a cart up an inclined ramp). Table 5-13 gives the model hopping features for all six questions with three groups of students.

In table 5-14, the questions is regrouped in two groups to calculate the f_s. The data under each question-shift describes the average number of model changes over all students. The data under "Pure" describes the percentage of students who never change their model on all questions in the group.

From table 5-13, we can see that the model-hopping among questions 8, 9, 11, 12 is significantly lower than if question 5 or 2 is involved. It reveals quite clear grouping of two question groups (2, 5) and (8, 9, 11, 12). This phenomenon can be seen more clearly in table 5-14 where the model hopping frequency between question 5 and 8 is significantly larger than all other question pairs indicating a major model shift between the two questions. For questions 8, 9, 11, and 12, the model hopping is much less frequent and there are 63% of students who used only one model to solve these four questions. As discussed earlier, the two groups of questions are based on two different story-lines. The results imply that the change of story-lines can often produce higher probability for students to use different models.

Table 5-13. FMCE model-hopping matrices

FMCE	2	5	8	9	11	12	f_h
Trd-pre	0.00	0.38	0.25	0.27	0.23	0.28	0.29
		0.00	0.42	0.47	0.43	0.45	
			0.00	0.18	0.18	0.23	
				0.00	0.22	0.19	
					0.00	0.16	
						0.00	
Trd-post	0.00	0.39	0.41	0.38	0.41	0.40	0.36
		0.00	0.54	0.54	0.49	0.44	
			0.00	0.20	0.28	0.28	
				0.00	0.32	0.17	
					0.00	0.22	
						0.00	
Rtp-post	0.00	0.25	0.42	0.36	0.42	0.36	0.35
		0.00	0.53	0.53	0.49	0.49	
			0.000	0.21	0.23	0.28	
				0.00	0.28	0.17	
					0.00	0.25	
						0.00	

Table 5-14. Student-based model-hopping of FMCE questions

FMCE	2→5	5→8	8→9	f_s	Pure
Trd-pre	0.38	0.42	0.18	0.33	0.47
Trd-post	0.39	0.54	0.20	0.38	0.33
Rtp-post	0.25	0.53	0.21	0.33	0.40
Average	0.34	0.49	0.20	0.34	0.40
	8→9	**9→11**	**11→12**		
Trd-pre	0.18	0.18	0.16	0.17	0.67
Trd-post	0.20	0.28	0.22	0.23	0.60
Rtp-post	0.21	0.23	0.25	0.23	0.62
Average	0.20	0.23	0.21	0.21	0.63

To further study this issue, let's look at the FCI data. The FCI test has five questions (5, 9, 18, 22, 28) on the FM concept and all these questions are "stand alone" questions each with an unique story-line and context settings. The results of the FCI questions are calculated in table 5-15 and table 5-16.

Table 5-15. FCI model hopping matrices

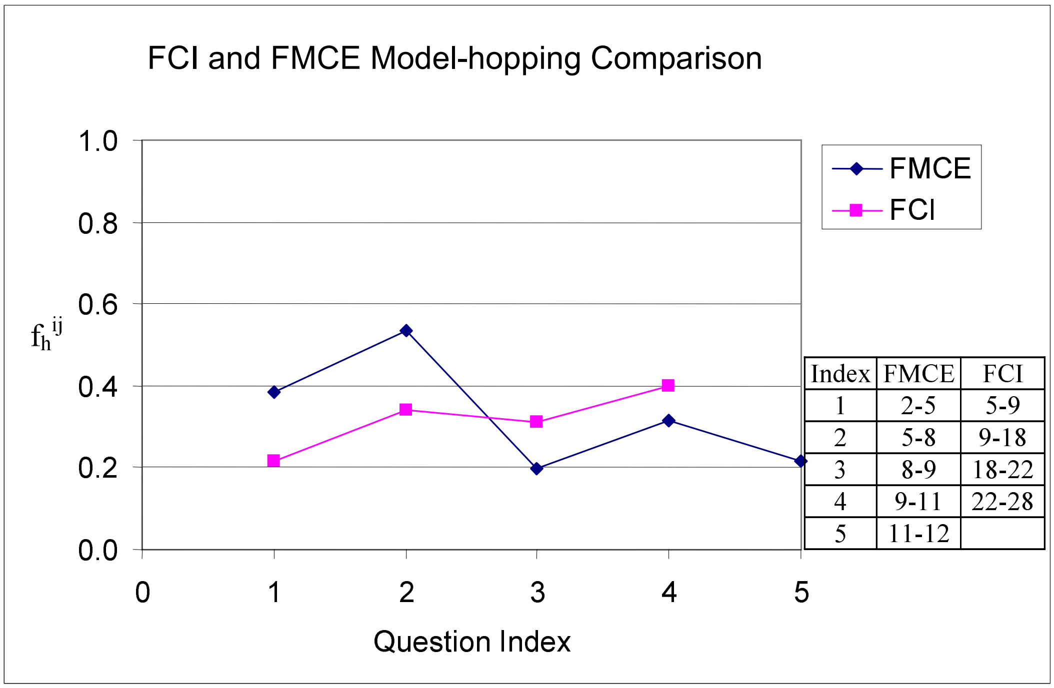
FCI	5	9	18	22	28	f_h
Trd-pre	0.00	0.22	0.32	0.23	0.40	0.32
		0.00	0.34	0.21	0.41	
			0.00	0.31	0.39	
				0.00	0.40	
					0.00	
Trd-post	0.00	0.27	0.39	0.24	0.38	0.34
		0.00	0.35	0.29	0.40	
			0.00	0.39	0.39	
				0.00	0.34	
					0.00	
Tut-post	0.00	0.19	0.45	0.22	0.47	0.36
		0.00	0.44	0.20	0.44	
			0.00	0.42	0.35	
				0.00	0.44	
					0.00	

Table 5-16. Student-based model-hopping for FCI questions

FCI	5→9	9→18	18→22	22→28	f_s	Pure
Trd-pre	0.22	0.34	0.31	0.40	0.32	0.35
Trd-post	0.27	0.35	0.39	0.34	0.34	0.34
Tut-post	0.19	0.44	0.42	0.44	0.37	0.33
Average	0.23	0.38	0.37	0.39	0.34	0.34

The results indicate that the FCI questions have more uniformly distributed model-hopping frequencies, and there is no strong indication on certain groupings of questions. On the other hand, the model-shift with the FMCE questions is mainly concentrated on the change between two question groups (see figure 5-9). As we can see from figure 5-9, there is a major change between FMCE question 5 and 8. This result agrees with the situation that all FCI questions have quite different contextual settings, which should generate comparatively large and more uniformly distributed probabilities for triggering different student models.

From the data in table 5-13 through table 5-16, we can see another interesting result. The model-hopping property is much less dependent on the backgrounds of the students than other evaluations such as model states. As indicated from both FCI and FMCE results, with the same set of questions in each test, the model hopping frequencies have quite consistent values for different students. This seems quite surprising since the difference between students can be very large and the result should inevitably be affected by student performances. For example, if a group of experts is given the test, we would expect have a very low model-hopping rate since they will almost always use an expert model in all questions. A similar situation will happen to a group of students with a very strong incorrect model except that this time the students are expected to use their incorrect model all the time. However, when the class of students gets into the mixed model stage, the random process in model triggering becomes a more important factor.



Index	FMCE	FCI
1	2-5	5-9
2	5-8	9-18
3	8-9	18-22
4	9-11	22-28
5	11-12	

Figure 5-9. FCI and FMCE model-hopping comparison

In addition, the model hopping only measures the frequency of student changing to use a different model and does not reflect the absolute use of a particular model. That is with four questions, student A can solve three with the expert model and one with an incorrect model whereas student B can solve three with an incorrect model and one with another incorrect model. Both student A and B may create a same f_s of 0.5 (depends on the order in which they use their models). But this f_s does not provide any information on what types of models the students are using. Such information is reflected in the student model states. Therefore, it is possible for students with different backgrounds but having a mixed model state to give similar model hopping behaviors.

When students are in mixed model states, the model-hopping frequency will have a significant contribution from the contextual structures of the questions, which are believed to be a major factor in model triggering. As a matter of fact, the data obtained from both tests shows strong dependence between f_h and the contextual structures of the questions. Therefore, we can use this measurement to evaluate the model triggering features of different instruments.

As indicated from the data, the two instruments (FCI and FMCE) have quite different features regarding model triggering. The FCI questions with diverse context settings allow higher and more uniformly distributed probabilities for triggering different models whereas the FMCE questions, which are often grouped with similar contextual settings, will produce smaller chances for students to be triggered into different models.

Model Triggering Probability

The model hopping frequency gives a differential evaluation on question-based model triggering features. It is also useful to find an absolute evaluation that gives the probability for different questions in a test to trigger a particular student model state.

Suppose there are a total of *m* questions in a test to measure a physics concept which has *w* physical models (including the null model). For a single student, we can construct a *w*-by-*m* student model data matrix, denoted as $\mathbf{A}^{k}$

$$\left[\mathbf{A}^{k}\right]=\left(\left|A_{1}^{k}\right\rangle,\ldots,\left|A_{i}^{k}\right\rangle,\ldots,\left|A_{m}^{k}\right\rangle\right) \qquad (5\text{-}13)$$

where $k = 1,\ldots, N$ represents different student and $\left|A_{i}^{k}\right\rangle$ is a *w*-dimension unitary vector representing the student model vector measured with the i^{th} questions. Depending on the design of the questions, $\left|A_{i}^{k}\right\rangle$ will have different structures. For a multiple-choice single-response type of question, we can only detect the triggering of a pure physical model (it is not impossible to design a question using one response to detect the existence of multiple physical models; however, it is often recommended not to do so because the interpretation may get too complicated). In this case, $\left|A_{i}^{k}\right\rangle$ will have one element equal to 1 and the rest equal to 0. With a multiple-choice multiple-response type of question, we can detect mixed state with a single question and $\left|A_{i}^{k}\right\rangle$ can have multiple non-zero elements.

In general, each question can have a non-zero probability to trigger any type of student model state. Therefore, we have to evaluate the model triggering probability with respect to a particular model state (we can map out all the possible state one by one and obtain a complete picture). Define $\left|B_{\eta}\right\rangle$ as the η^{th} model state that is to be studied. Using $\left|B_{\eta}\right\rangle$ as the template, we can obtain a question-based model triggering probability vector, denoted as $\left|T_{\eta}^{k'}\right\rangle$. Then we can write:

$$\left|T_{\eta}^{k'}\right\rangle=\left(\left\langle B_{\eta}\middle|A_{1}^{k}\right\rangle,\ldots,\left\langle B_{\eta}\middle|A_{i}^{k}\right\rangle,\ldots,\left\langle B_{\eta}\middle|A_{m}^{k}\right\rangle\right)^{T}=\left[A^{k}\right]^{T}\left|B_{\eta}\right\rangle \qquad (5\text{-}14)$$

To normalize $\left|T_{\eta}^{k'}\right\rangle$, we can create a new vector $\left|T_{\eta}^{k}\right\rangle$ by inserting an additional dimension containing the complements of all elements in $\left|T_{\eta}^{k'}\right\rangle$:

$$\left|T_\eta^k\right\rangle = \frac{1}{\sqrt{m}}\left(\left\langle B_\eta \middle| A_1^k\right\rangle, \ldots, \left\langle B_\eta \middle| A_i^k\right\rangle, \ldots, \left\langle B_\eta \middle| A_m^k\right\rangle, \sqrt{m - \sum_{j=1}^{m}\left\langle B_\eta \middle| A_j^k\right\rangle^2}\right)^T \quad (5\text{-}15)$$

It is easy to see that $\left|T_\eta^k\right\rangle$ is normalized to 1:

$$\left\langle T_\eta^k \middle| T_\eta^k\right\rangle = \frac{1}{m}\left(\left\langle B_\eta \middle| A_1^k\right\rangle^2 + \ldots + \left\langle B_\eta \middle| A_i^k\right\rangle^2 + \ldots + \left\langle B_\eta \middle| A_m^k\right\rangle^2 + m - \sum_{j=1}^{m}\left\langle B_\eta \middle| A_j^k\right\rangle^2\right) = 1$$

Depending on the goals of the analysis, we can perform two different calculations to obtain $\left|T_\eta^k\right\rangle$. If we want an exact match, then we can redefine the dot product to be

$$\left\langle B_\eta \middle| A_i^k\right\rangle = \begin{cases} 1 & \text{if } B_\eta = A_i^k \\ 0 & \text{if } B_\eta \neq A_i^k \end{cases}$$

This calculation separates different model states even when they are overlapping (non-orthogonal). Therefore, the results give an exclusive clustering of student model states. If we consider the similarities between model states, then the ordinary dot product is used in calculation and the results produce a non-exclusive clustering (fussy clustering) of student model states. In the case where each question can only detect a pure physical model, both methods will give the same results.

Using $\left|T_\eta^k\right\rangle$, we can construct a question-based model triggering probability density matrix, denoted as $\mathcal{T}_\eta$. $\mathcal{T}_\eta$ is an *m*-by-*m* matrix representing the projection of the overall model triggering probability on a particular model state. We can write:

$$\mathcal{T}_\eta = \frac{1}{N}\sum_{k=1}^{N}\left|T_\eta^k\right\rangle\left\langle T_\eta^k\right|$$

With a *w*-dimension model space, if each question can only detect a pure physical model state, then $\eta = 1, \ldots, w$. In this case only the projections on pure physical models can be studied. If mixed model states can be measured with a single questions (e.g. multiple-choice multiple-response type), then η can be larger than w and the projections on mixed model states can also be studied.

Using $\mathcal{T}_\eta$, we can study the model triggering probabilities for the different questions on the η^{th} model state by performing eigenvalue decomposition or cluster analysis (see chapter 4 for discussion on eigenvalue decomposition and cluster analysis). As an example, using FCI data (UMd students) the question-based model triggering probability in the projection on the correct model is calculated in table 5-17.

Table 5-17. Question-based model triggering probability on the projection of the correct model

	Trd-Pre	Trd-Post		Tut-Post	
σ^2	0.81	0.64	0.24	0.67	0.21
5	0.03	0.28	0.81	0.78	0.33
9	0.02	0.36	0.81	0.86	0.31
18	0.14	0.21	0.64	0.43	0.41
22	0.01	0.26	0.81	0.84	0.3
28	0.13	0.17	0.66	0.38	0.44

As indicated from analysis in chapter 4, the initial states for both tutorial classes and traditional classes are very similar. Therefore, table 5-17 only shows the initial results with traditional classes. Since the eigenvalue for the initial state is quite large, only the primary state is listed. The initial state indicates that all five questions have very low probability to trigger students to use the correct model and question 18 and 28 have a slightly higher probability. The post instruction results of the traditional classes show a dominant state with quite low probabilities for all questions to trigger the correct model. The dominant state of the tutorial classes has significantly higher probabilities to trigger a correct model. Quite interestingly, for both states, question 18 and 28 show smaller probabilities compare to the other questions. This result is consistent with the model-hopping evaluation where model-hopping frequencies have a relatively larger value when one of the two questions is involved.

Model Triggering Consistency

If we are interested in the overall consistency of model states triggered by different questions regardless of the types of models, we can construct a model triggering correlation matrix and perform factor analysis to study this issue. Define the question-based model triggering correlation matrix for the kth student with $\mathbf{R}_t^k$. Using Eq. (5-13), we can write:

$$\left[\mathbf{R}_t^k\right]=\left[\mathbf{A}^k\right]^T\cdot\left[\mathbf{A}^k\right]=\begin{bmatrix}\left\langle A_1^k\middle|A_1^k\right\rangle & \left\langle A_1^k\middle|A_2^k\right\rangle & \cdots & \left\langle A_1^k\middle|A_m^k\right\rangle\\ \left\langle A_2^k\middle|A_1^k\right\rangle & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ \left\langle A_m^k\middle|A_1^k\right\rangle & \cdots & \cdots & \left\langle A_m^k\middle|A_m^k\right\rangle\end{bmatrix}$$
$$=\begin{bmatrix}1 & r_{12}^k & \cdots & r_{1m}^k\\ r_{12}^k & 1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \vdots\\ r_{1m}^k & \cdots & \cdots & 1\end{bmatrix} \qquad (5\text{-}16)$$

Then the question-based model triggering correlation matrix for the whole class, denoted with $\mathbf{R}_t$, can be find with:

$$[\mathbf{R}_t] = \frac{1}{N}\sum_{k=1}^{N}\mathbf{R}_t^k \qquad (5\text{-}17)$$

The creation of this correlation matrix is different from the method used in traditional factor analysis. The method used here is called model-based correlation construction where the correlation of two model states is calculated with the similarity between the two states (a dot product between two unitary vectors).[8]

Using $\mathbf{R}_t$, we can perform factor analysis to study if there exists any general patterns on triggering similar models with different questions. As an example, the question-based model consistency correlation matrix calculated in table 5-18 with post instruction FCI data (tutorial classes at UMd).

Table 5-18. Question-based model consistency correlation matrix with post tutorial FCI data (UMd students)

	5	9	18	22	28
5	1	0.81	0.55	0.78	0.53
9	0.81	1	0.56	0.8	0.56
18	0.55	0.56	1	0.58	0.65
22	0.78	0.8	0.58	1	0.56
28	0.53	0.56	0.65	0.56	1

	5	9	18	22	28
5	1	0.81	0.55	0.78	0.53
9	0.81	1	0.56	0.8	0.56
18	0.55	0.56	1	0.58	0.65
22	0.78	0.8	0.58	1	0.56
28	0.53	0.56	0.65	0.56	1

> 0.7

	5	9	18	22	28
5		0.81		0.78	
9	0.81			0.8	
18					
22	0.78	0.8			
28					

< 0.7

	5	9	18	22	28
5			0.55		0.53
9			0.56		0.56
18	0.55	0.56		0.58	0.65
22			0.58		0.56
28	0.53	0.56	0.65	0.56	

Figure 5-10. Binary relation of the model consistency correlation matrix for threshold 0.7

From table 5-18, we can see that the model states triggered with questions 5, 9, and 22 have high correlation (model states are similar). The model states triggered with question

18 and 28 are much less similar. This relation can be shown more clearly by decomposing the correlation matrix with a threshold (see figure 5-10).[9] These results show clear grouping of questions 5, 9, and 22 based on the similarities of the model states triggered by these questions. The model states triggered by question 18 and 28 have relatively large differences compared to the other three questions. Although between the two questions the correlation is slightly higher, the absolute value is still small (~0.65). This indicates that the two questions also trigger different model states. Combining the results from the model triggering probability, it implies that question 18 and 28 has less probability to trigger a correct model but students are often affected by one of the two questions at a time, not both. These results are consistent with the results from model hopping and question-based model triggering probability.

As we can see, the different methods give different aspects of the model triggering features of the question with different emphasis. In order to get a complete picture, we need to actively use all the methods and combine the results to obtain appropriate interpretations.

Structures of Student Models

Student models are context dependent. The contextual information is strongly involved in the mental processes of model construction, model triggering and model application. To understand student model structures and model operation, it is important to systematically study the context dependence of student models.

Physical Features

Physical contexts exist in various forms and contain numerous different features. For a particular physics concept domain, through systematic qualitative and quantitative research we can identify a finite set of relevant features, defined as *physical features*. A physical feature describes a unique contextual aspect of a physics representation and is considered relevant to the physics concept of interests by experts and students. In general, physical features have the following characteristics:

1. The definition of physical features for a particular physics concept domain is population dependent.
   A physics context can involve many different features and can also be related many different physics concepts. With different physics concepts, different sets of physical features can be defined for the same context and the definition is dependent on the background of the student population. For example, when considering Newton III, experts know that under any circumstances, the force during an interaction is opposite and has the same magnitude. On the other hand, to a group of naïve students, they often consider that velocity, mass, acceleration and source of the force will make differences to the force during an interaction. To overly naïve students (e.g. kindergarten students), it is also possible for these students to consider the size of the object as an important factor. Therefore, when defining physical features, we need to take into account the student background and identify these physical features through systematic research.

2. The interpretation of a physical feature may be different for different population. For a particular physical feature, different people may have different interpretations although they may be using the same terminology. This can also be a reason for many student difficulties where they are using the same words as experts would use but constructing a totally different picture than what experts would expect. In defining physical features, the recognized interpretations from our science community are used. In developing diagnostic instruments, we need to consider such misinterpretation and design appropriate tools to measure this information.

Using physical features, we can study the fine details of the structures of student models for a particular physics concept. We can also study the dynamical process of student model operation with respects to the different physical features. For example, we can study the details of model triggering features to see if different physical features may have similar or different contributions to the triggering of a particular student model state. We can also study student model-evolution process to see if students have similar or different model-change patterns with different physical features.

Examples of Physical Features and Current Instruments

The two concept domains, Force-Motion and Newton III, in mechanics have quite different structures on physical features. With Force-Motion, there is only one physical feature involved – velocity. Students often think that there is a force in the direction of motion and the magnitude of the force is related to the velocity. In this case, the design of diagnostic instruments and the analysis of student data are quite straightforward.

On the other hand, the Newton III is far more complicated than Force-Motion. From research, we can identify four different physical features: velocity (V), mass (M), acceleration (A) and source of the force (or who is pushing, denoted with P).[10] In the literature, researchers often characterize the common incorrect student model in Newton III with the so-called dominant agent model, where students often think that during interaction, a dominant agent will exert a larger force.[11] The dominant agent can be a particular physical feature such as mass, velocity, the source of the force, etc., or it can be a combination of several different physical features. However, in the literature, there is not adequate research on how the individual physical features or the combination of them may contribute to the student model structures and model operations. As discussed earlier, the questions in both the FCI and FMCE tests are not designed with clear isolation of these different physical features. Therefore, in previous discussions with FCI and FMCE tests, the different type of student models with emphasis on different physical features were collapsed into a single general dominant-agent model. The detail of how individual physical features may contribute to student models is considered as the fine structures of the general model.[12]

Many of the questions in current instruments are not designed with isolated physical features. For example, the question shown in figure 5-11 mixes two physical features, mass and pushing, together. If a student answers that the big guy exerts a larger force, there is no further evidence to tell if the incorrect response is generated based on considerations on the mass or on the pushing.

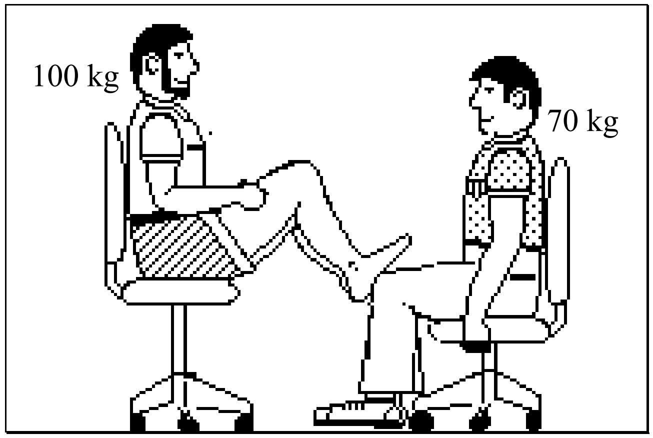


Figure 5-11. Sample question on Newton III with mixed physical features of mass and pushing

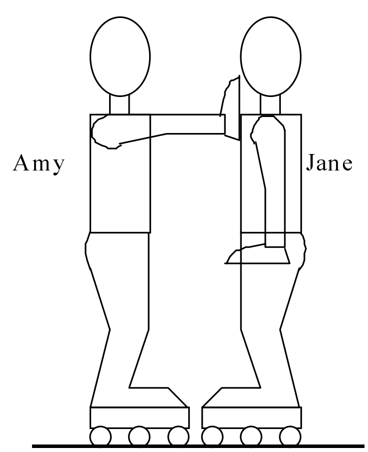
A New Multiple-Choice Instrument with Isolated Physical Features

In order to probe the details of student models with individual physical features, we need to design appropriate instrument with isolated physical features. As an example, in this section I introduce a new multiple-choice instrument for Newton III where each question only deals with one physical feature. In this instrument, for each of the four physical features three questions with different context settings are designed. Figure 10-12 shows two examples, one on velocity and one on pushing. The complete test is included in Appendix C.

Velocity

The small truck has the **same weight** as the car does. At the time of collision, both vehicles travel at a constant speed but the small truck is moving at a **slower** speed than the car.

Pushing



They both have a **same mass** of 50 kg. Amy then suddenly pushes outward with her hand, causing both to move. In this situation, while Amy's hands are in contact with Jane, which choice describes the forces?
…

Figure 5-12. Newly designed questions on Newton III with isolated physical features

Mathematical Modeling and Algorithms

- **Microstate Model Evaluation**

For each physical feature, we can define a set of physical models and construct a model sub-space. Represent the physical models with M_{η}^{l} where l represents the different physical features and η identifies the different physical models for the l^{th} phsyical feature. Suppose we have a total of L physical features involved. Then $l = 1, \ldots, L$. With the example of Newton III, there are four physical features ($L = 4$) and we can label them as V (l=1), M (l=2), P (l=3), and A (l=4). For each of the physical features, we can construct a 3-dimension model sub-space (In other cases, it is possible that the numbers of physical models with different physical features can be different). Using the physical feature of pushing (P) as an example, we can define the following physical models:

M_0^3: The null model

M_1^3: The force has the same magnitude and opposite direction during the interaction regardless the source of the force. (correct expert model)

M_2^3: The party exerting the force will exert a larger force during interaction. (incorrect student model)

With a set of questions designed with a particular physical feature, we can measure the probability for a single student to use the different physical models. Then the single student model state for the k^{th} student in a class can be represented with (see chapter 2 and chapter 4 for details on definitions of single student model states):

$$\left|u_k^l\right\rangle = \left(\sqrt{q_{0l}^k}, \sqrt{q_{1l}^k}, \sqrt{q_{2l}^k}\right)^T, \quad l = 1, \ldots, L \qquad (5\text{-}18)$$

and

$$\left\langle u_k^l \middle| u_k^l \right\rangle = \sum_{\eta=0}^{w-1} q_{\eta l}^k = 1$$

Since we are looking at the sub-space of a general model structure, the model state with a particular physical feature is called a *microstate*. When only one physical features is involved, e.g. the Force-Motion, the microstate becomes the same as the general model state defined in previous chapters.

Using the single student microstate, we can construct class microstate model density matrix with a particular physical features. Denote this matrix with $\mathcal{D}_l$. We can write:

$$\mathcal{D}_l = \frac{1}{N}\sum_{k=1}^{N}\left|u_k^l\right\rangle\left\langle u_k^l\right|$$

Using $\mathcal{D}_l$, we can perform eigenstate analysis or clustering analysis in a similar manner described in chapter 4.

- **Macrostate Model Evaluation**

When the overall student model structure with all the involved physical features is of the interests, we can construct model states representing different macroscopic features of student model structure. These model states are called *macrostates*. For simplicity, it is assumed that the microstates all have the same dimension. Then putting together all the microstates of a single student on different physical features, we can construct a *w*-by-*L* student model data matrix in a similar manner as in Eq. (5-13). Thus we can write:

$$[\mathbf{U}^k] = \left(\left| u_k^1 \right\rangle, \ldots, \left| u_k^l \right\rangle, \ldots, \left| u_k^L \right\rangle \right) \tag{5-19}$$

Similarly, we can study the projection of the model data matrix on a particular model state defined with a template model state $\left| V_\mu \right\rangle$. A specific projection is defined as a macrostate which represents the contributions (in probabilities) from the different physical features to a particular model state (we can perform two different projections – exclusive and non-exclusive as discussed previously). Denote the μ^{th} macrostate of the k^{th} student with $\left| F_\mu^k \right\rangle$. We can write:

$$\left| F_\mu^k \right\rangle = \frac{1}{\sqrt{L}} \left(\left\langle V_\mu \middle| u_1^k \right\rangle, \ldots, \left\langle V_\mu \middle| u_l^k \right\rangle, \ldots, \left\langle V_\mu \middle| u_L^k \right\rangle, \sqrt{L - \sum_{l=1}^{L} \left\langle V_\mu \middle| A_l^k \right\rangle^2} \right)^T \tag{5-20}$$

In cases where the dimensions of the microstates are not the same, we need to do additional treatment depending on the goals of the analysis, e.g. reorganizing or manually selecting model elements in a microstate to extract relevant information.

Using the macrostates, we can also construct class macrostate model density matrix denoted as $\mathcal{F}_\mu$, which can be obtained with:

$$\mathcal{F}_\mu = \frac{1}{N} \sum_{k=1}^{N} \left| F_\mu^k \right\rangle \left\langle F_\mu^k \right|$$

Similarly we can perform eigenstate analysis and clustering analysis to study the unique aspects of the student models with different physical features.

- **Microstate Model Consistency**

Sometimes, it is useful to look at the effects from different physical features on the complete microstate rather than a particular projection. In this case, we can study the consistency of the microstates with the different physical features. Using similar formulations in Eq. (5-16), we can construct the single student microstate correlation matrix $\mathbf{R}_F^k$:

$$\left[\mathbf{R}_F^k\right]=\left[\mathbf{U}^k\right]^T\cdot\left[\mathbf{U}^k\right]=\begin{bmatrix} 1 & \langle u_1^k | u_2^k\rangle & \cdots & \langle u_1^k | u_l^k\rangle \\ \langle u_2^k | u_1^k\rangle & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \langle u_L^k | u_1^k\rangle & \cdots & \cdots & 1 \end{bmatrix} \tag{5-21}$$

The class microstate correlation matrix, $\mathbf{R}_F$, can be obtained with:

$$\left[\mathbf{R}_F\right]=\frac{1}{N}\sum_{k=1}^{N}\mathbf{R}_F^k$$

Similarly, we can perform factor analysis to study if there exist any general patterns for microstates with different physical features.

Analysis Results

- **The Students**

The data used in this analysis is collected at the Kansas State University with the newly developed multiple-choice test on Newton III. Currently, only pretest data is available for analysis. The students in this study are from five different courses described in table 5-19.

Table 5-19. Backgrounds of students from Kansas State University (traditional instruction)

Courses	Types of Courses	Majors	Math Pre-requisites
Physical World	Algebra, Mech.	Liberal arts	No math
General Physics 1	Algebra, Mech.	Life science	Algebra
General Physics 2	Algebra, E&M	Life science	Algebra
Engineering Physics 1	Calculus, Mech.	Eng and Phys	Calculus
Engineering Physics 2	Calculus, E&M	Eng and Phys	Calculus

- **Microstate Analysis**

The student class microstates with the four physical features of Newton III are calculated and plotted in figure 5-13. Currently, only the pre-instructional data is available, so the model states plotted in figure 5-13 represent the initial states for each class. However, from the description of the student population (table 5-19), general physics 2 (GP2) and engineering physics 2 (EP2) are the second courses in the introductory series and mechanics is the topic of the first courses (general physics 1 and engineering physics 1). Therefore, we can use the data from GP2 and EP2 as the data of post traditional instruction (approximately).

From figure 5-13, we can see that for the physical features of mass and velocity, students all have a dominant consistent incorrect model. The popularity of the incorrect model decreases somewhat as the level of courses gets higher – from 90% (GP1) to 60% (EP2). The secondary model states of GP2 and EP2 are also plotted which indicates a consistent correct model. As we can see that for all classes student models are very consistent, i.e. very few mixed states. In this case, students in a class can be partitioned in two groups, one with a consistent incorrect model and the other with a consistent correct model. The eigenvalue of the corresponding model state vector gives an estimation of the size of the group.

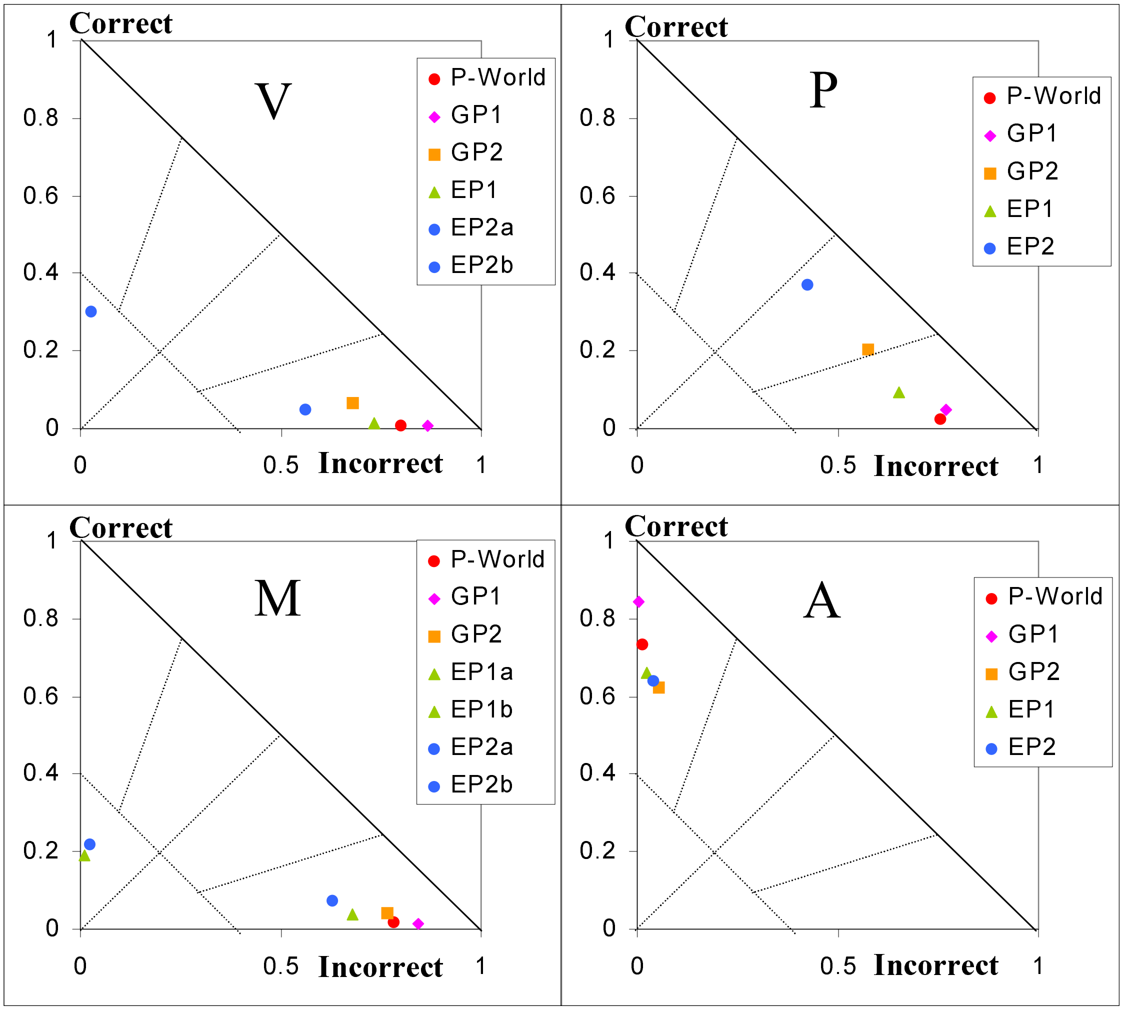


Figure 5-13. Student class microstates for Newton III with different physical features

Student model states with the physical feature of acceleration show the opposite to the situations with mass and velocity. In this case, most students hold a consistent "correct" model where they consider acceleration irrelevant. Student model states are also quite consistent (little mixing). Although students give correct responses on the related question, it does not mean that student models are the same as the expert model. As indicated from interviews, the reason for students to consider acceleration irrelevant is not that they truly understand the nature of Newton III but rather that they consider the velocity is the major factor and acceleration is something related to velocity and will not make direct effect. Quite interestingly, the results indicate that more students in higher level classes change their ideas on this issue.

With the physical feature of pushing, student model states show completely different patterns. The low level classes still show a dominant consistent incorrect model. As the level of students gets higher, student model states become more mixed. The most advanced class (EP2) has nearly a perfectly mixed model state with a quite large eigenvalue (~0.8) which indicates that most students have similar model states. This is very different from the situations with the other physical features and implies a different concept-change process.

- **Macrostate Analysis**

Using Eq. (5-20), the macrostates created with non-exclusive protection on incorrect student models are calculated and the probability plot is graphed in figure 5-14. Since for all classes, the eigenvalues for the primary macrostates are around 0.9, using these states alone is enough to show the classes' behavior.

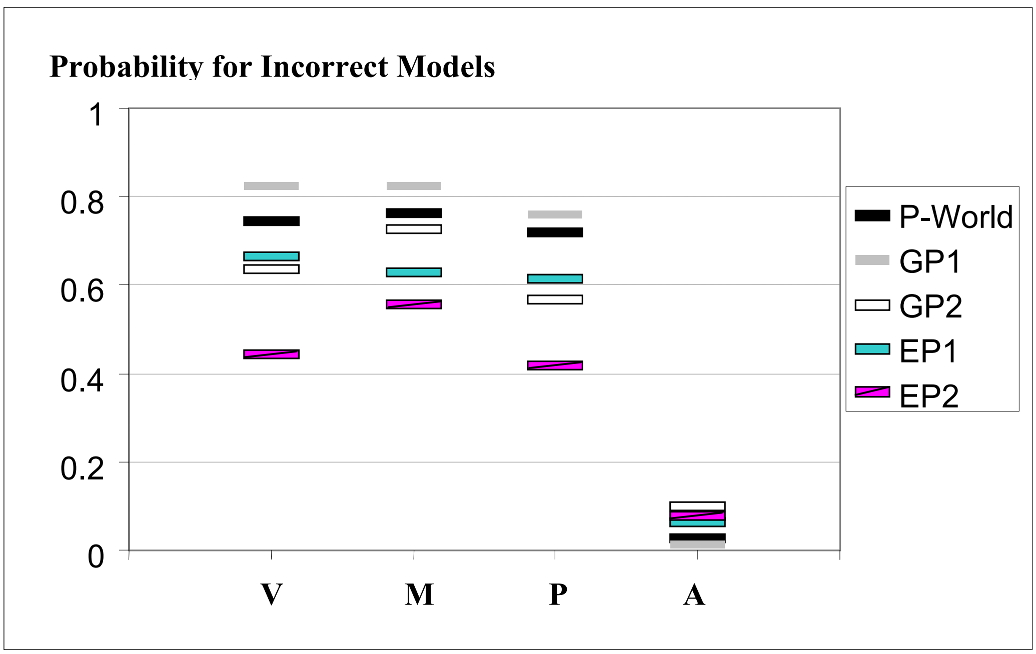
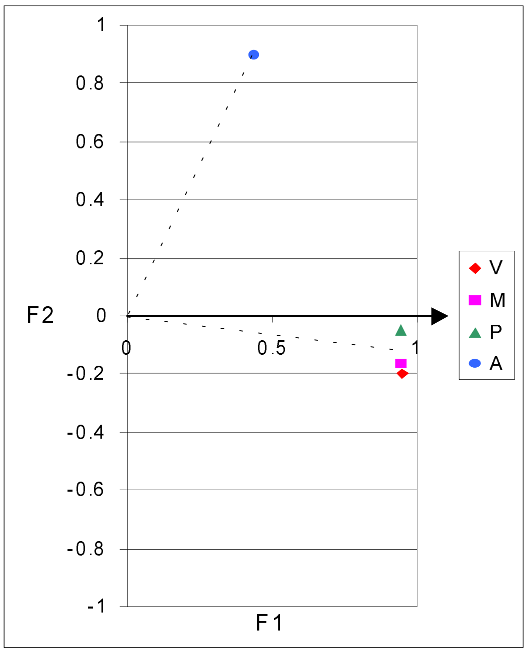


Figure 5-14. Probabilities for students to use the incorrect models with different physical features. The macrostate is calculated based on a similarity projection (non-exclusive). The probability is calculated with the primary macrostates (for all classes, the eigenvalues for the primary macrostates are around 0.9).

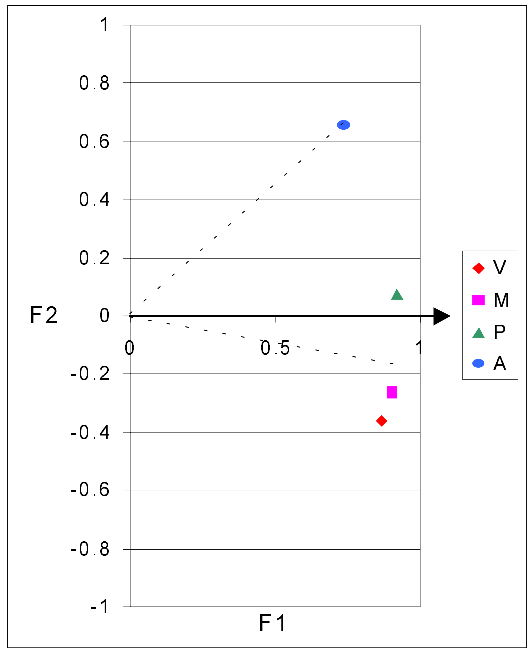
As we can see, for the classes (P-world, GP1, EP1) without instruction on mechanics, the four physical features all have similar contributions to the incorrect models. After instruction with mechanics, the physical feature of mass has relatively large probabilities to trigger incorrect models. This result is consistent with the model plot shown in figure 5-13. Notice that although the overall contribution from the physical feature of pushing is similar to that from velocity, from the model plot in figure 5-13, we know that the contribution with velocity is from a consistent incorrect model and the contribution with pushing is from a component of a mixed model state. This type of information can also be extracted with macrostate when using exclusive projection.

- **Model Consistency**

Using the method discussed in Eq. (5-21), we can evaluate the consistency of the student microstates with different physical features. To analyze this consistency, we can use model-based factor analysis.[13] In figure 5-15, the results for two classes are plotted.



General Physics 1



Engineering Physics 2

Figure 5-15. Model consistency factors

From figure 5-15, in the low level class (without instruction on mechanics), the three physical features (V, M and P) all trigger similar student microstates and the physical feature of acceleration (A) triggers a different (nearly orthogonal) student microstate. In the advanced class after instruction with mechanics, we can see that the V and M still stay close but the P has significant changes. This result is consistent with the model plot shown in figure 5-13.

Student Model Evolution

From the different analysis, we can see that the students have different model structures with the different physical features. Student model changes also follow different paths with the different physical features. In the context of Newton III, the results indicate that after traditional instruction students often change to mixed model states with the physical feature of pushing. As discussed by many researchers, the stage of mixed model states is often an important intermediate step for a complete favorable concept change.[14]

To further study student model evolution process, qualitative research is important. As indicated from interviews (10 students from KSU in the level of P-world). The initial states of these students are the same as students in P-world. After instruction, 80% of the students still have strong consistent model with V, M. But, almost every one of them start to challenge their reasonings with P. However, most of them are still in an uncertain mixed state. On acceleration, most (80%) of them think it irrelevant and they always pull out the V issue as reasoning.

When asked for the reasoning on the physical features of pushing, most students specifically quoted that "when you pushing something, you get pushed back". Many of them even repeat the sentence that "the force is equal and opposite" and they tried to use this idea to reason through the problems. In interviews, several observations are quite common to most students:

1. They often use the two sentence in the questions where pushing is the major issue. Otherwise, they use M or V instantly without even bother to recall the two sentences, which all of them can memorize, especially for the first one, which they often give examples with personal experience. It seems that the two sentences were tied with pushing only.
2. When they use the two sentences, first one is no a problem, but they still have problems with the second one and have the tendency to think the pusher exert larger force. So they can give contradictory answers on similar questions resulting in a mixed model state.
3. It appears when multiple physical features are involved at the same time, students can get confused (more undecided). They have a tendency to favor the "stronger physical features" which are often V and M.
4. An embedded general "root" for all these problems is that students often take the results (effect) as the causes in reasoning.

With the results from the qualitative and quantitative methods, we can infer a possible explanation for the fact that the model changes in Newton III starts with the physical feature of pushing: It appears that "Pushing" is often the most common example used to introduce Newton III and students can also easily learn the experience of being pushed back. Therefore, students can have significant model changes on this physical feature even with traditional instruction. On the contrary, student model changes with other physical features are fairly insignificant.

As a short summary, from the study of this example, student models show different structures with different physical features and the student model evolution also show different patterns with different physical features. Such information is often unavailable when using instruments designed with mixed physical features. As an example, the new instrument and algorithms are found useful in measuring and analyzing the details of the structures of student models. With this new method we can obtain detailed quantitative information of student models with a particular physical features as well as the macroscopic aspects of student models with different physical features. In addition, the results from these evaluations can provide direct insightful information on student understandings for both research and instruction.

Implications on Instrument Design

Model Measurement with Multiple-choice Questions

As discussed in chapter 4 and 5, there are in general two ways to model student responses on multiple-choice questions, the item-based modeling (IbM) and the CPA. With item-based modeling, all choices of multiple-choice questions should have one-to-one or many-to-one choice-to-model correspondence. With CPA, some of the questions in a group can have one-to-many choice-to-model correspondence, however, a large overlap of the response patterns often increases the uncertainty of measured model states.

The choices of the questions need to be carefully designed and validated through research. Usually, we want the choices to be straightforward to assure a high successful rate for students to correctly apply their models triggered by the questions. It is also possible to design a sequence of questions with similar context settings but different levels of complexity to obtain an evaluation on the ability of students applying their models correctly. Such instruments need to be supported by extensive qualitative research to ensure correct interpretations of the results.

Resolution of Model Measurement

The resolution of model measurement is primarily determined by the number of questions. In order to detect the mixing of two physical models, it requires a minimum of two MCSR type questions. Using MCMR type questions, it is possible to detect model mixing with one question. To evaluate class model states, the number of questions for each physics concept is recommended to be larger than the number of physical models involved (when multiple physical features are involved, for each physical feature, we need to make the number of questions be larger than the number of physical models with that physical feature). To obtain an accurate measurement of a single student model state, the number of questions needs to be significantly larger than the number of physical models. See Eq. (2-5) and figure 4-10 for details of the uncertainty relation in model measurement.

Fine Structure of Models: Isolation of physical features

When student models are associated with a variety of physical features, e.g., the Newton III, we can design questions to isolate the different physical features and to obtain measurement on the finer details of student models with a particular physical feature. Instruments with isolated features allow us to see different student model structures with different physical features and different model evolution processes with different physical features. In general, when multiple physical features exist, we need to conduct appropriate qualitative research and design instruments that allow the measurement of student models with the different physical features or the combination of them.

Context Settings of Questions

As indicated from the analysis results, specific context structures of questions can make significant contributions to the student model triggering processes. Therefore, in order to obtain robust measurement of student model states, we need to include questions with diverse context settings. This not only can provide more stable measurement on mixed model states, but also can reduce possible bias in the assessment where students may be doing well on one type of context settings and still have a lot of problems with some other contexts related to the same concept.

The Cyclic Process to Develop Model-Based Diagnostic Instrument

As widely recognized in the PER community, the development of a research-based diagnostic instrument need to be integrated in the cyclic process of research, instruction and development. With model analysis, we can design model-based diagnostic instruments. The methodologies and algorithms in model analysis not only can be used to obtain quantitative evaluations of student understandings but also can be used in the process of developing model-based multiple-choice instruments to make quantitative assessment of various features of a test. Figure 5-16 is a schematic diagram that shows how the different methods and tools in model analysis can be applied in the process of developing a model-based diagnostic instrument.

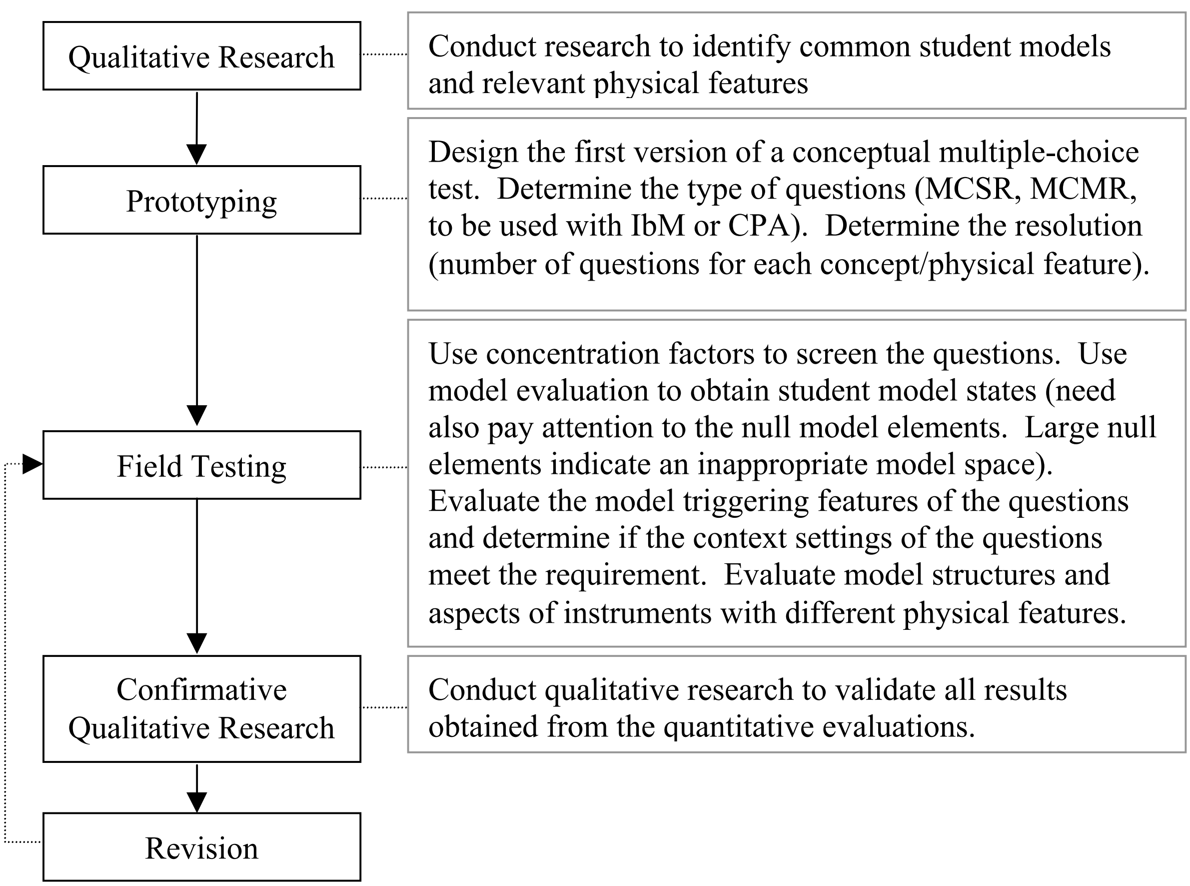


Figure 5-16. The cyclic process of model-based multiple-choice instrument design

Summary

In this chapter, I have dsicussed several examples that require additional algorithms for data processing. The first example is based on a multiple-choice multiple-response question used to study student understanding of mechanical waves. Students are found to have a mixed initial model state where they treat waves as particles and as waves using both types of models inconsistently. The students in tutorial classes are found to move to reasonably consistent correct model after instruction while students with traditional instruction still struggle with mixed models.

The second example is the FMCE test, which deals with similar concepts as the FCI does. However, many FMCE questions often involve multiple physical features and the different physical models often have overlapping responses.   Thus these questions cannot be modeled with response from a single question. In responding to such situation, I developed a new modeling scheme – the CPA, which uses the coherence of the student responses on a series of questions. With the CPA, student models are studied and compared with the results from FCI. In addition, the model triggering features of the two tests are also studied with several new measurements, the model-hopping frequency, question-based model triggering probability, and question-based model consistency evaluation. The FCI questions, with more diverse context settings, are found to have a more uniformly distributed probabilities for triggering different student models.

As a further development of model analysis, student model structures and model evolution process are studied based on physical features. Based on this idea, an multiple-choice instrument on Newton III with isolated physical features is developed. Several new algorithms are also developed to do data analysis.

References and Endnotes:

[1] M. Wittmann, "Making sense of how students come to an understanding of physics: An example from mechanical waves," Ph.D. dissertation, University of Maryland, 1998.

[2] See reference 1.

[3] The *Tools for Scientific Thinking* (TST) *Force and Motion Conceptual Evaluation* is available from the Center for Science and Mathematics Teaching, 4 Colby St., Tufts University, Medford, MA 02155.

[4] R. K. Thornton, "Enhancing and evaluating students' learning of motion concepts." Chapter in *Intelligent Learning Environments and Knowledge Acquisition in Physics*, A. Tiberghien and H. Mandl, eds. (Berlin-Heidelberg-New York, Springer Verlag, NATO Science Series, 1992)

[5] Data collected by L. Hadley L, UMd.

[6] 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

[7] This explains the result observed by Thornton (private communication) that FCI and FMCE scores correlate well except on the low end, where FMCE grades are found near zero, but few FCI grades are lower than 20%.

[8] Please contact the author for more details on model-based correlation construction and model-based factor analysis.

[9] This method is similar to the binary relation used in cluster analysis. E.g. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall.

[10] A significant part of qualitative research (student interviews) is conducted at the Kansas State University. In literature, there are also a lot of published results that indicate the existence of the 4 physical features.

[11] I. Halloun and D. Hestenes, "The initial knowledge state of college physics students," *Am. J. Phys.* **53** (11), 1043-1055 (1985).

[12] L. Bao, et. al., "Model Analysis on Fine Structures of Student Models," AAPT Announcer, Jan. 2000

[13] Model-based factor analysis is different from traditional factor analysis. This new method is developed based on the Model Analysis theoretical framework. Due to time limitation, the details of this algorithm is not included in this thesis. Please contact the author for any current write-ups on this issue.

[14] See reference 6