Appendix B: Supplemental Calculations and Simulations for the Concentration Factor

The Level Coding for Concentration Factor

Let us begin with a two-peak situation where most of the responses are concentrated on two choices. Table B-1 shows the simulated data for a range of possible combinations and their corresponding ***C*** factor values. We can see that the value of ***C*** ranges between 0.23 to 0.36 when 80% to 90% of the responses are concentrated on two choices. Table B-2 shows the simulated data with no dominant peaks. We can see that when the responses are somewhat evenly distributed among three choices, the value of ***C*** becomes small. Even with a 90% concentration among three of the choices, the largest ***C*** is smaller than 0.17. Table B-3 shows the case when the majority of the responses concentrated on one choice. At different ratios between 60% and 90%, the values of ***C*** range from 0.4 to 0.8.

Simulated data for quantization

Table B-1. Two peaks situation (N=100)

80% concentrated on two choices						90% concentrated on two choices					
A	B	C	D	E	***C***	A	B	C	D	E	***C***
40	40	8	6	6	0.23	40	50	5	3	2	0.354
41	39	8	6	6	0.231	41	49	5	3	2	0.352
42	38	8	6	6	0.231	42	48	5	3	2	0.35
43	37	8	6	6	0.233	43	47	5	3	2	0.348
44	36	8	6	6	0.235	44	46	5	3	2	0.347
45	35	8	6	6	0.238	45	45	5	3	2	0.347
46	34	8	6	6	0.241	46	44	5	3	2	0.347
47	33	8	6	6	0.246	47	43	5	3	2	0.348
48	32	8	6	6	0.25	48	42	5	3	2	0.35
49	31	8	6	6	0.255	49	41	5	3	2	0.352
50	30	8	6	6	0.261	50	40	5	3	2	0.354

Table B-2. Data more widely distributed among three or more choices (N=100)

80% concentrated on three choices						90% concentrated on three choices					
A	B	C	D	E	***C***	A	B	C	D	E	***C***
20	27	33	10	10	0.062	20	30	40	5	5	0.169
21	27	32	10	10	0.057	21	30	39	5	5	0.163
22	27	31	10	10	0.054	22	30	38	5	5	0.157
23	27	30	10	10	0.051	23	30	37	5	5	0.152
24	27	29	10	10	0.048	24	30	36	5	5	0.148
25	27	28	10	10	0.047	25	30	35	5	5	0.144
26	27	27	10	10	0.046	26	30	34	5	5	0.141
27	27	26	10	10	0.046	27	30	33	5	5	0.138
28	27	25	10	10	0.047	28	30	32	5	5	0.137
29	27	24	10	10	0.048	29	30	31	5	5	0.136
30	27	23	10	10	0.051	30	30	30	5	5	0.135

Table B-3. One peak situation (N=100)

Variable concentrations on one choice					
A	B	C	D	E	***C***
60	30	5	3	2	0.409
65	25	5	3	2	0.455
70	20	5	3	2	0.512
75	15	5	3	2	0.579
80	10	5	3	2	0.653
85	5	5	3	2	0.735
90	0	5	3	2	0.823

The Maximum Line of the State Density Plot

Let's define the problem as the followings: Suppose we have 100 responses on one question with 5 choices (N = 100 and m = 5 with choices of a, b, c, d, and e). Without loosing generality, we will always make "e" as the correct answer. Thus the score, S, will represent the number of students who have chosen "e" (here the score can be from 0 to 100). Since the value of S is already determined, now the problem becomes one dimension less with a new constraint of:

$$a + b + c + d = N - S$$

Then it is all up to the configuration of the "a", "b", "c" and "d" to determine the concentration factor on the S-***C*** plot. For each value of the score, the maximum density of states happens when the number of the possible combinations is at the greatest. For the 4-element case, let's take "a" as the variable analogous to the score in the 5-item case. For a given value of "a", the largest number of possible combinations occurs when the value range of all the rest three are equal and at their largest possible values which indicates that a = 0. Therefore the maximum line is at where a = 0 and the rest of the three are approximately equal to each other. This is equivalent to the low boundary of the 5-element S-***C*** plot with one choice always being zero. In figure B-1, the calculated maximum line is plotted together with the state density.

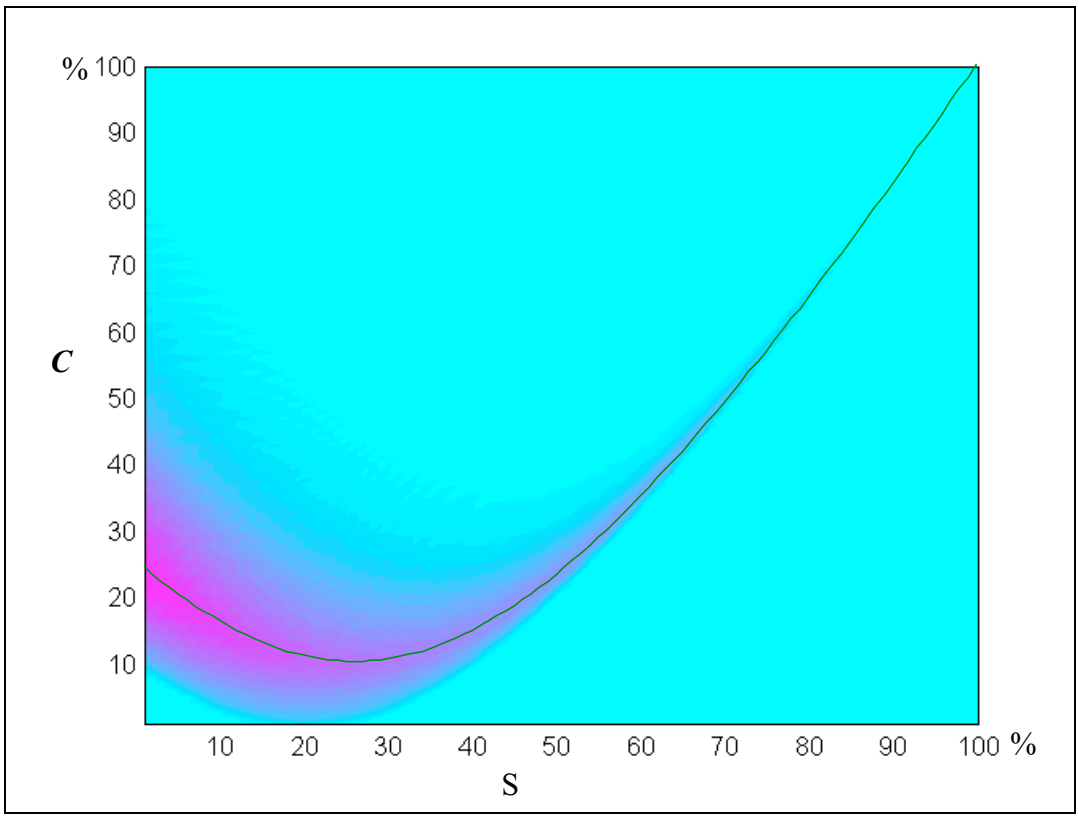

Figure B-1. The peak density line of the S-***C*** state density attractor

Numerical Evaluations

1. Numerical Measures of the States and Shift Vectors

The S-***C*** plot provides us a graphical representation of the data. In many cases, quantitative evaluations are also desired. Therefore we designed the following numerical measures that can give quantitative evaluations of various aspects of the student performance.

a. The Agreement

Since the student data is now represented with points and vectors in a two-dimensional graph, the performance needs to be evaluated with the agreement between the vectors representing student responses and the vectors for the favorable responses.

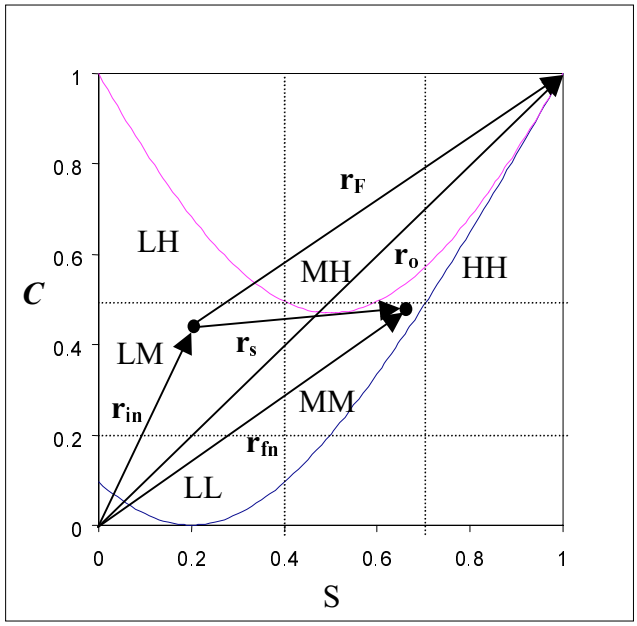

Figure B-2. Vector analysis of S-***C*** state shifts

As in figure B-2, we can construct a vector from the origin to the initial point, denoted as $\mathbf{r}_{in}$, to represent the initial state. The vector, $\mathbf{r}_{fn}$, representing the final state can be defined similarly. We also know that the most favorable response is at point (1,1). Therefore we can always define the optimum response vector, $\mathbf{r}_o$, as one starting from the origin and ending at point (1,1). To see how good the $\mathbf{r}_{in}$ and $\mathbf{r}_{fn}$ agree with the optimum state, we can define the agreement measure as:

$$A_{in} = \frac{\mathbf{r}_{in} \cdot \mathbf{r}_o}{|\mathbf{r}_o|^2} \qquad \text{and} \qquad A_{fn} = \frac{\mathbf{r}_{fn} \cdot \mathbf{r}_o}{|\mathbf{r}_o|^2}$$

It is obvious that the value of the agreement is bounded between 0 and 1.

b. The Possible Agreement Improvement (PAI)

Once we have the initial state agreement (A_{in}) and the final state agreement (A_{fn}), the improvement of the agreement can be evaluated. In a similar manner as the definition of the possible gain on scores,[1] we define the possible agreement improvement as:

$$PAI = \frac{A_{fn} - A_{in}}{1 - A_{in}}$$

c. The Possible Favorable Shift

The student improvement can be represented with a vector, $\mathbf{r}_S$, starting from the initial point towards the final point. Then for each initial state, there exists a most favorable shift vector, $\mathbf{r}_F$. Obviously $\mathbf{r}_F$ is a vector starting from the initial point to point (1, 1). In reality,

the student actual shift vector will often be different from the most favorable one. To measure how good their actual shift is, we can again find the agreement, denoted as "s", between the actual shift vector and the most favorable shift vector:

$$s = \frac{\mathbf{r}_S \cdot \mathbf{r}_F}{\left|\mathbf{r}_F\right|^2}$$

As we can see, if only the score component of the shift vectors is evaluated, it becomes 1-D and it will give the same calculation as "h" – the scaled improvement in score defined by Hake.[2] There are possibilities for the absolute value of "s" to be greater than 1 if the initial point is close to "HH" with a large unfavorable shift towards "LH" or "LL". In many cases when the initial score and concentration have similar values, the measure of "s" and "PAI" gives similar results. The details are explained later.

d. The Possible Concentration Gain

Since the concentration factor tells important information about the student mental models, it is useful to know the improvement on this issue. The possible concentration gain is defined as

$$d = \frac{\boldsymbol{C}_{\mathrm{fn}} - \boldsymbol{C}_{\mathrm{in}}}{1 - \boldsymbol{C}_{\mathrm{in}}}$$

where $\boldsymbol{C}_{\mathrm{in}}$ (and $\boldsymbol{C}_{\mathrm{fn}}$) represents the initial (and final) concentration factor.

The "PAI" and the "s"

Suppose we have an initial point at (S1, D1) and a final point at (S2, D2), then

$$A_{\mathrm{in}} = \frac{S1 + D1}{2}, \quad A_{\mathrm{fn}} = \frac{S2 + D2}{2} \quad \text{and} \quad PAI = \frac{A_{\mathrm{fn}} - A_{\mathrm{in}}}{1 - A_{\mathrm{in}}} = \frac{(S2 - S1) + (D2 - D1)}{(1 - S1) + (1 - D1)}$$

The possible favorable shift can be calculated as:

$$s = \frac{(S2 - S1)(1 - S1) + (D2 - D1)(1 - D1)}{(1 - S1)^2 + (1 - D1)^2} = \frac{(S2 - S1) + (D2 - D1)\dfrac{(1 - D1)}{(1 - S1)}}{(1 - S1) + (1 - D1)\dfrac{(1 - D1)}{(1 - S1)}}$$

Then it is easy to see that when S1 and D1 have similar values, the s and PAI are about the same, which is what we have got from our real data analysis.

Transformation between Γ and *C*

The regions for different response types shown in Figure 3-3 have a different configuration in the S-Γ plot, which is mapped out in Figure B-3.

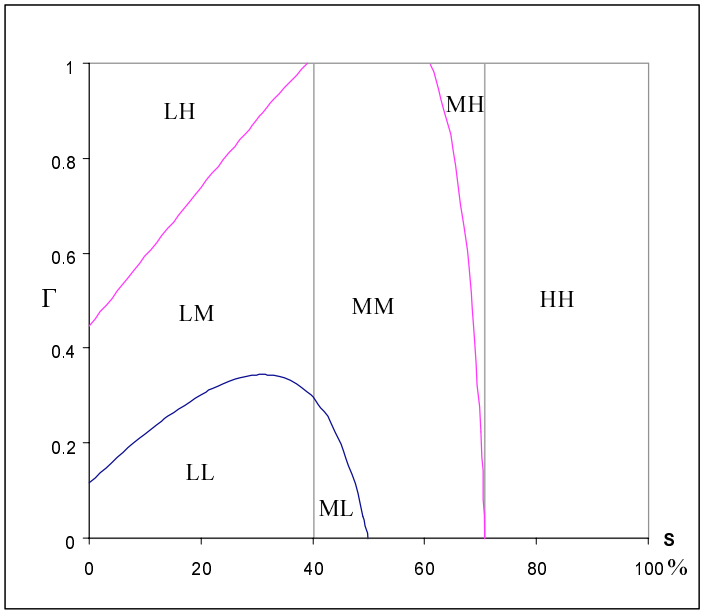

Figure B-3. Regions of different response types in S-Γ plot

Since Γ represents the variation part of the overall concentration which can also be obtained with another transformation by using

$$\boldsymbol{C}_{\mathrm{V}} = \frac{\boldsymbol{C} - \boldsymbol{C}_{\min}}{\boldsymbol{C}_{\mathrm{Max}} - \boldsymbol{C}_{\mathrm{Min}}} \qquad \text{(B-1)}$$

It is important to know how the two calculations are different. In figure B-4, the three constant $\boldsymbol{C}_{\mathrm{V}}$ lines are plotted on the S-Γ domain where we can see that the two transformations are different especially when the score is large. This is because the effect of score on the concentration is getting larger when score is large.

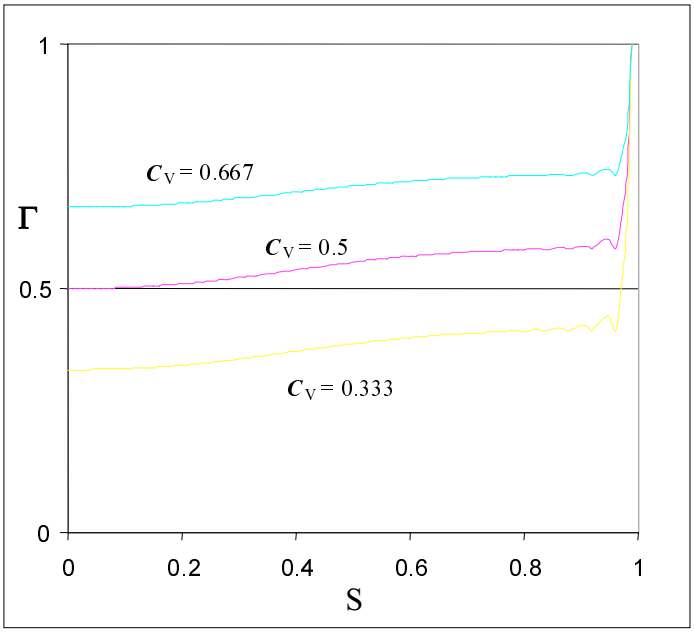

Figure B-4. S-Γ at constant C_V

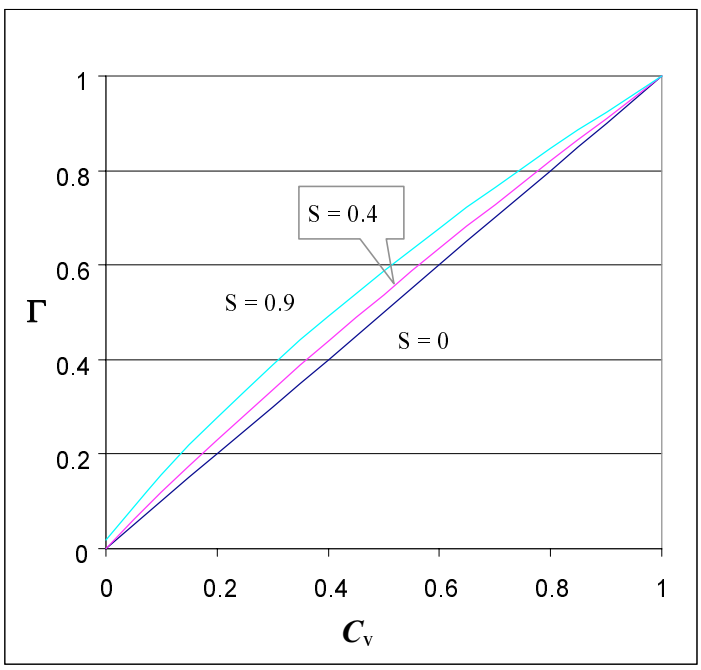

Figure B-5. Γ vs. C_V at different score

To see more clearly about the relation between Γ and C_V, we graphed the Γ vs. C_V in figure B-5 with different scores. When score is zero, since there is no contribution from the score the two transformation give the same results. In general the two transformations give similar results (off by 10% when score is at 50%) when the score is not too large. But

the physical meaning of Eq. (3-10) is more straightforward and Eq. (B-1) has a singular point when score equals 100%. Therefore, in our analysis, Eq. (3-10) is used.

Additional Comments on Linearization of Γ

In cases when there is only one major distractor, the student behavior on the uncommon distracters is very close to a random situation. That is student reponses on the unpopular distracters are like random guessing. Therefore, the lower boundary line, which is obtained by assuming that the responses on the rest of the choices are evenly distributed, can be a good approximation of the response curve describing the relation between Γ and the largest single choice concentration. With this assumption, the square root of Γ follows nicely with the actual concentration on the major distractor (shown in table B-4).

Table B-4. Concentration Deviation when the score is 50%

Concentration on One	Effective Concentration	Relative Concentration	Average Γ	$\sqrt{\Gamma}$
40%	15%	38%	0.15	0.39
60%	35%	58%	0.35	0.59
75%	50%	67%	0.55	0.74

References:

[1] Richard R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66**, 64-74 (1998).

[2] See reference 1.