

# Mathematical Features of Concentration Analysis

Lei Bao

*Department of Physics, The Ohio State University,  
174 W 18<sup>th</sup> Ave., Columbus, OH 43210*

## Abstract

In this paper, we introduce some mathematical features of the concentration factor, a method to extract the information of the conditions of student mental models by analyzing the distributions of students' responses on multiple-choice questions. We discuss the state density of the concentration factor, the random attractor of *SC* states, and the transformations between the concentration factor and the concentration deviation. The results can provide useful information for researchers to apply concentration analysis to study student data.

## I. INTRODUCTION

Multiple-choice concept surveys are useful instruments to assess student understanding in large classes.<sup>1</sup> However, traditional analysis often relies on scores and doesn't provide information on how students respond with incorrect answers, which contain a large amount of valuable information. In our research, we developed a new method, *concentration analysis*, to measure the distribution of students' responses on multiple-choice questions.<sup>2</sup> The results of concentration analysis can be used to study if the students have common incorrect models or if the test is effective in measuring student models.

Concentration analysis provides an analytical method for analyzing the concentration/diversity of students' responses on a particular multiple-choice question. A key element of this method is the *concentration factor*, which is a function that maps the response of a class on a multiple-choice question to the interval  $[0, 1]$  with zero corresponding to students selecting a random distribution of answers and one corresponding to all students selecting the same answer.<sup>3</sup> The details of applying concentration analysis to study student responses are discussed extensively in the references.<sup>4</sup> In this paper, we introduce some interesting features of concentration factor, which can provide useful information for interpreting the results from using this mathematical tool.

We begin the paper in section II with a brief review of the definition of concentration factor. In section III and IV, we discuss the state density and random attractor of the *SC* plot. In section V, we discuss the transformation between the concentration factor and the concentration deviation – concentration of the incorrect responses. We conclude with a summary.

## II. THE CONCENTRATION FACTOR

As indicated from research on student learning,<sup>5</sup> students' responses to problems can often be considered as the result of their applying a small number of mental models.<sup>6</sup> If the choices of a multiple-choice question is designed to reflect the popular mental models that the students have, the class' responses on this question will be concentrated on the choices that are associated with the models common to the population. On the other hand, if the students have no consistent models of the topic, their responses can be close to a random distribution among all the choices. Therefore, the way in which the students' responses are distributed can yield information on the situations of students' models.

To conveniently measure the distribution of a class' responses on one multiple-choice question, we define the *concentration factor*,  $C$ , as a function of students' responses that takes a value in  $[0, 1]$  with 1 representing perfectly concentrated responses and 0 representing uniformly distributed responses.<sup>7</sup>

Suppose we give a multiple-choice question with  $m$  choices to  $N$  students. The concentration factor can be written as

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

where  $n_i$  is the number of students in the class who selected choice  $i$ . With  $N$  students, we have

$$\sum_i^m n_i = N \quad (2)$$

Using concentration factor, we can evaluate the conditions of students' use of their models. For example, with  $m = 5$  (most FCI questions have 5 choices),  $C > 0.5$  often indicates that more than 70% of the students have selected the same choice (one concentration peak).<sup>8</sup> When  $C$  has a value between 0.2 and 0.5, usually two choices often attracts more than 80% of the students' responses (two concentration peaks). When  $C$  is smaller than 0.2, students responses are close to a uniform distribution (no concentration peaks), which often indicates that students don't have dominant models or that students respond to the question by random guessing.

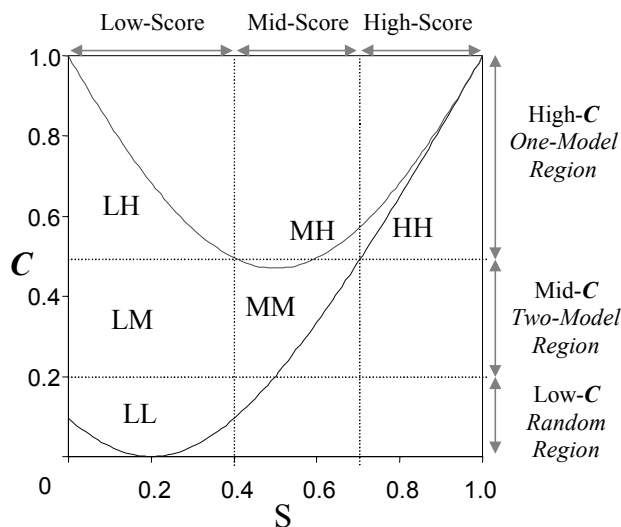


Figure 1. Due to the constraint between the score and concentration factor, data points can only exist in the area between the two boundary lines.

Combining the score ( $S$ ) with the concentration factor, we can analyze whether the question triggers a common “misconception”. A situation of low score ( $S < 0.4$ ) but high concentration value ( $C > 0.5$ ) is represented with an LH type of response and often indicates that students are likely to have a very popular incorrect model. In a situation of medium score ( $0.4 \sim 0.7$ ) and medium concentration ( $0.2 \sim 0.5$ ), referred as an MM type of response, students are often in a mixed state between of the correct and incorrect models.<sup>9</sup> These results can be conveniently represented with an  $SC$  plot using the score and the concentration factor as two axes. With this representation, the students' responses on each question can be represented as a point on the  $SC$  plot. Due to the entanglement between the score and the concentration factor (see eq. (2)), data points can only exist in certain regions on an  $SC$  plot. With a  $m$ -choice question, this allowed region are the area between the two boundaries describe with:

$$C_{MIN}(S) = \frac{\sqrt{m}}{\sqrt{m}-1} \times \left( \sqrt{(m-1) \left( \frac{1-S}{(m-1)} \right)^2 + S^2} - \frac{1}{\sqrt{m}} \right) \quad (3)$$

and

$$C_{MAX}(S) = \frac{\sqrt{m}}{\sqrt{m}-1} \times (\sqrt{(1-S)^2 + S^2} - \frac{1}{\sqrt{m}}) \quad (4)$$

As an example, using eqs. (3) and (4) with  $m = 5$ , the boundary of the allowed region is plotted in Figure 1. The regions of student response types and implications on possible conditions of student models are also marked out.<sup>10</sup>

### III. THE STATE DENSITY OF THE CONCENTRATION FACTOR

Since the number of the total responses of a class is usually large, the number for all the possible combinations of these responses can be very large. Defining each possible combination as an *SC* state, with  $N = 100$  and  $m = 5$  the number of all possible states can be calculated with equation (5)

$$\mathcal{L} = \sum_{n_1=0}^{100} \left\{ \sum_{n_2=0}^{100-n_1} \left[ \sum_{n_3=0}^{100-n_1-n_2} \left( \sum_{n_4}^{100-n_1-n_2-n_3} 1 \right) \right] \right\} = 4598126 \cong 4.6 \times 10^7 \quad (5)$$

where  $n_i$  represents the number of students who select the  $i^{\text{th}}$  choice.

An *SC* state can be represented with a vector  $\vec{R}$  :

$$\vec{R} = (n_1, n_2, \dots, n_i, \dots, n_m) \quad \text{where} \quad \sum_{i=1}^m n_i = N \quad (6)$$

From eq (1), we can see that redistributing the components (without changing the values) in  $\vec{R}$  can produce the same concentration value. Therefore, a point on an *SC* plot can be associated with multiple *SC* states, which can be described with a state density function defined as  $\rho(S, C)$ . Obviously,  $\rho(S, C)$  satisfies

$$\int_{S=0}^1 \int_{C=0}^1 \rho(S, C) \cdot dSdC = 1 \quad (7)$$

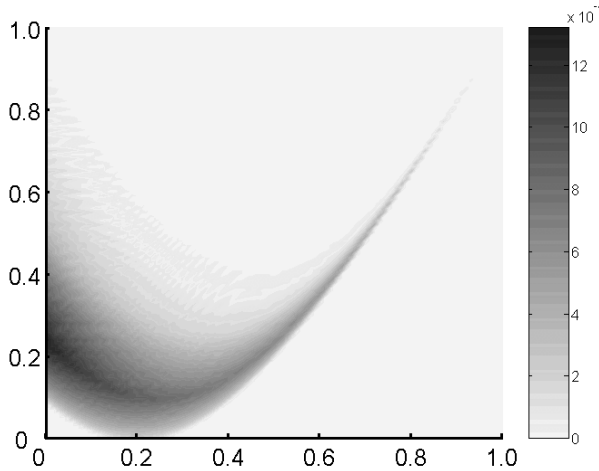


Figure 2. *SC* state density of an *SC* plot.

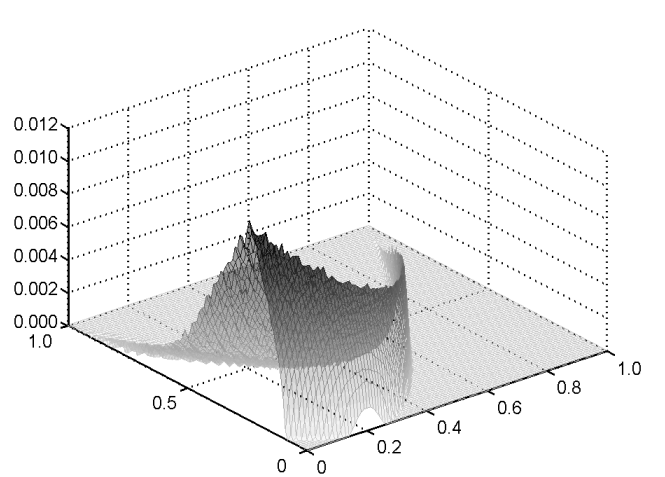


Figure 3. Three-D surface plot of *SC* state density.

It is interesting to see how the  $SC$  states are distributed on an  $SC$  plot. Assuming each state is equally probable and using an exhaustive enumeration method (with  $N = 100$  and  $m = 5$ ), we calculated and plotted the state density in Figure 2. The value of the density is obtained by first calculate the number of  $SC$  states in a rectangular area of  $\Delta S = 0.01$  and  $\Delta C = 0.01$  and then normalize this number with  $\mathcal{L}$ . It can be shown that the peak line, which goes across the points for largest state densities at given scores, is the low boundary of the  $SC$  plot of the same  $m$  and  $N$  values under the constraint that one of the incorrect choices is never selected.<sup>11</sup>

For clearer representation of the details of the state density, Figure 3 shows a three-dimensional surface plot of the  $SC$  states in the space spanned by  $S$ ,  $C$  and  $\rho$ .

#### IV. THE RANDOM ATTRACTOR OF THE CONCENTRATION FACTOR

In practice, the probability for each  $SC$  state to occur is affected by the students and the questions; and therefore cannot be an uniform value. As a special case, if we assume all the responses generated by the students are based on random guessing, it is then possible to simulate the  $SC$  state density for random responses. Using a Monte Carlo method by assuming each individual student is independent and respond randomly, we did a computer simulation for the process of giving a 5-choice question to a class of 100 students for 1 million times and obtained a random attractor. The density plot is shown in Figure 4. The three-dimensional surface plot of this random attractor is also shown in Figure 5. The value of the density is logarithmic to show more detail of the low-density area.<sup>12</sup>

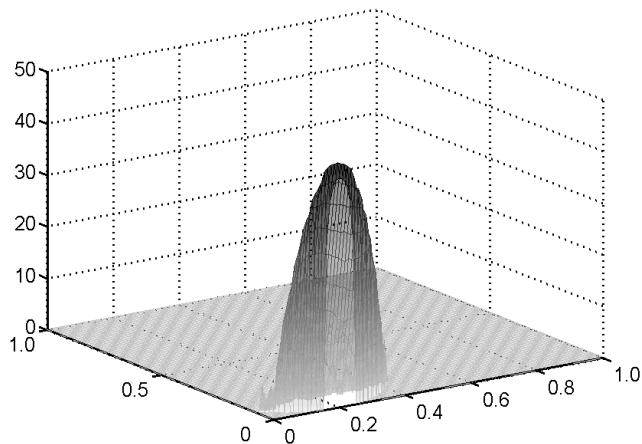
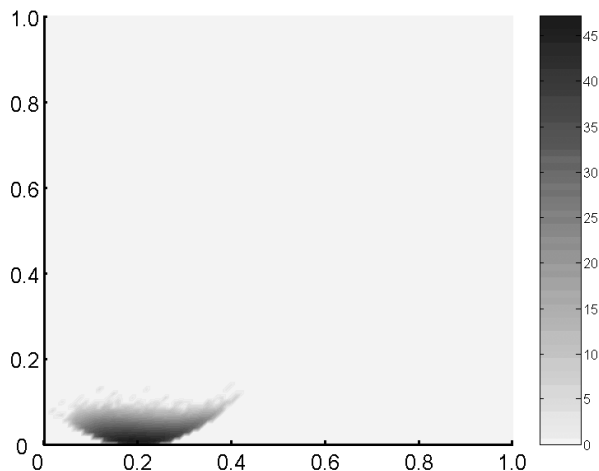


Figure 4. Random attractor on an  $SC$  plot. The plotted state density reflects the result of  $10\log(\mathcal{L}\times\rho+1)$ .

Figure 5. Three-D surface plot of the random attractor.

As expected, the attractor is concentrated around the minimum point of the  $S-C$  lower boundary ( $S = 0.2$ ,  $C = 0$ ) with  $\Delta S = \pm 0.1$  and  $\Delta C = 0.1$ , which is at the center of the low score and low concentration region (see the LL region in Figure 1). This result confirms our assumption for the LL type of response being an indication for random guessing.<sup>13</sup> Further, since the dynamic range of this random attractor is very large, about 50 dB (5 orders of magnitude). This implies that if most students in a class are guessing randomly, it is very unlikely for the class' responses to form an  $SC$  state outside the random attractor.

To see more clearly the position of the peak of the random attractor, the 3-D surface in Figure 5 is projected on to the plane spanned by  $C$  and  $\rho$  (see Figure 6). We can see a very sharp peak around  $S = 0.2$  and  $C = 0.01$ . Note that the peak is not at  $C = 0$ , which corresponds to a uniformly distributed responses ( $n_i = N/m$ ), the expectation value of a random guessing ensemble. This is because that to make

$C$  very close to 0, the randomly generated responses need to be in a near-perfect uniform distribution among the five choices. This is less likely than if some of the choices have small deviations from the expectation value.<sup>14</sup> The peak center being at  $S = 0.2$  and  $C = 0.01$  reflects such deviations in the order of 10% around the perfect uniform distribution for the four incorrect choices.<sup>15</sup>

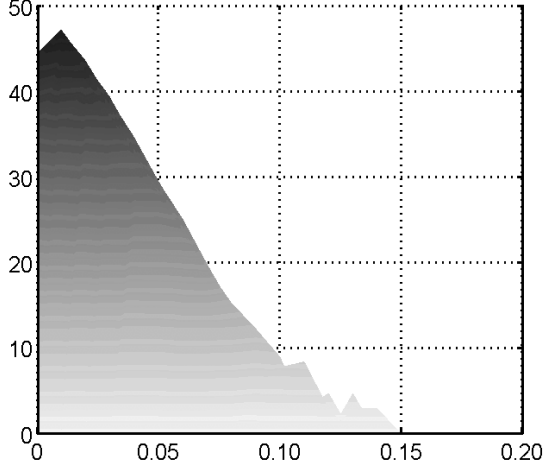


Figure 6. A projection of the 3-D random attractor on the  $C$ - $\rho$  plane at  $S = 0.2$ . It shows that the peak occurs around  $C = 0.01$ .

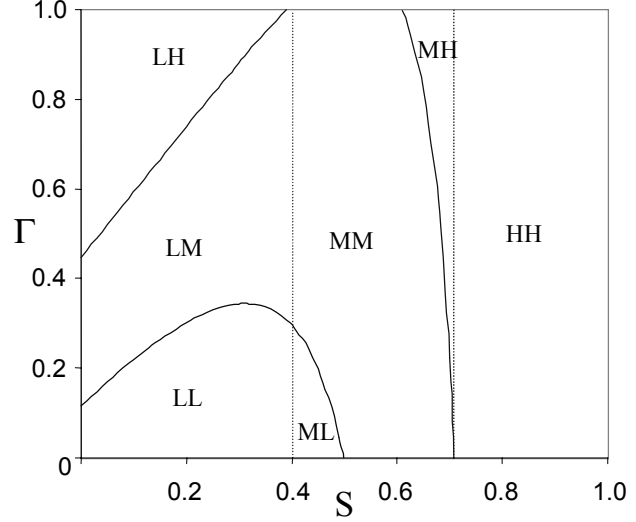


Figure 7. Transformation between an  $SC$  plot and an  $S\Gamma$  plot.

## V. THE CONCENTRATION DEVIATION

The concentration factor gives the overall concentration of student responses. Due to the constraint of eq. (2), the  $C$  given by eq. (1) is dependent on the score and the data points on an  $SC$  plot are restricted in a strangely shaped attractor. The entanglement between score and the concentration factor can limit the information from concentration analysis at large score where the concentration is determined mostly by the score. In order to disentangle the concentration from the score and to see the detail of the distribution of the incorrect responses, we designed a new variable, the *concentration deviation* –  $\Gamma$ , which gives the concentration of the incorrect responses:

$$\Gamma = \frac{\sqrt{m-1}}{\sqrt{m-1}-1} \times \left( \frac{\sqrt{\sum_{i=1}^m n_i^2 - (S \cdot N)^2}}{(N - S \cdot N)} - \frac{1}{\sqrt{m-1}} \right) \quad (8)$$

Since  $\Gamma$  and  $S$  are independent,  $\Gamma$  can have any value within the full range of  $[0, 1]$ . We can also construct an  $S\Gamma$  plot to study the details of the incorrect responses, and there is no restriction on the plotting area. All the  $SC$  states on an  $SC$  plot can be mapped onto an  $S\Gamma$  plot. Figure 6 is an  $S\Gamma$  plot that shows the different response type regions of an  $SC$  plot (see Figure 1). The two curved lines correspond to the horizontal lines (representing constant  $C$ ) in an  $SC$  plot, which separate the response type regions. Note that the allowed region on an  $SC$  plot is mapped to the whole area on an  $S\Gamma$  plot; therefore the top and low boundaries of the allowed region of an  $SC$  plot correspond to the two horizontal lines for  $\Gamma = 1$  and  $\Gamma = 0$  in the  $S\Gamma$  plot.

From eqs. (3) and (4), we can see that a particular value of score determines the range (absolute boundary) of the possible values of concentration factor. The variation of  $C$  within the boundary is

determined by the distribution of the incorrect student responses. This suggests us to consider another variable, denoted with  $C_V$ , which gives the scaled value of the variation part of the overall concentration:

$$C_V(S) = \frac{C(S) - C_{\text{Min}}(S)}{C_{\text{Max}}(S) - C_{\text{Min}}(S)} \quad (9)$$

It seems that  $\Gamma$  and  $C_V$  give somewhat similar measures; therefore it is useful to know the relation between the two functions. Figure 8 shows the  $S$ - $\Gamma$  relation at three constant  $C_V$  values. As we can see, the two calculations are different and the difference increases as the score gets larger.<sup>16</sup>

In Figure 9, we also graphed  $\Gamma$  vs.  $C_V$  with different scores. This result indicates that in general the two calculations are not too different (off by 10% when score is at 90%). Therefore, when score is not too large, we can use  $\Gamma$  as an approximation for the scaled variation portion of the overall concentration.

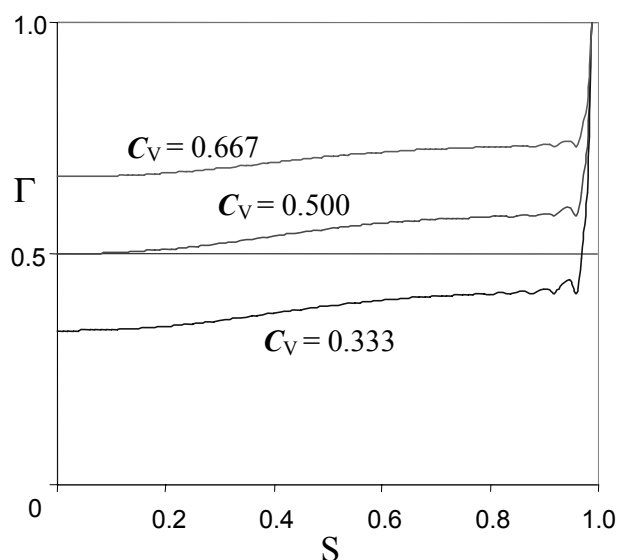


Figure 8.  $S$ - $\Gamma$  relation at constant  $C_V$ .

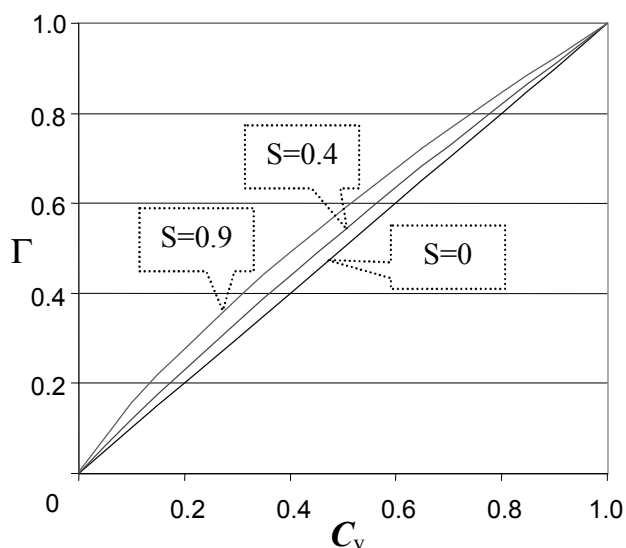


Figure 9.  $\Gamma$ - $C_V$  relation at constant score.

## VI. SUMMARY

In this paper, we introduced some mathematical features of the concentration analysis. We discussed the state density of the concentration factor, the random attractor of the  $SC$  plot, and the transformations between the concentration deviation and the concentration factor. The features of the random attractor indicate that if the students in a class all respond randomly, the  $SC$  states will most like be inside the random attractor centered around  $S = 0.2$  and  $C = 0.01$ , with  $\Delta S = \pm 0.1$  and  $\Delta C = 0.1$ . The properties of the concentration deviation reveal that when score is less than 90% the concentration deviation has similar values to the scaled variation part of the overall concentration.

## ACKNOWLEDGEMENTS

The author would like to thank professor Edward F. Redish and professor Lenard Joessem for their constructive discussions.

## Endnotes and Reference

---

- <sup>1</sup> D. Hestenes, M. Wells, and G. Swackhammer, "Force Concept Inventory", *The Physics Teacher*, **30**, 141-153 (1992).
- <sup>2</sup> L. Bao and E. F. Redish, "Concentration Analysis: A Quantitative Assessment of Student States," Accepted for publication in *PERS of Am. J. Phys.*, July 2001.
- <sup>3</sup> L. Bao, "Dynamics of Student Modeling: A Theory, Algorithms, and Application to Quantum Mechanics," Ph.D. dissertation, University of Maryland, December 1999. Available on request from [www.physics.ohio-state.edu/~lbao](http://www.physics.ohio-state.edu/~lbao).
- <sup>4</sup> References 1 and 2.
- <sup>5</sup> L. Bao and E. F. Redish, "Model Analysis: Assessing the Dynamics of Student Learning," submitted to *Cognition and Instruction*;  
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;  
J. Minstrell, "Facets of students' knowledge and relevant instruction", In: *Research in Physics Learning: Theoretical Issues and Empirical Studies*, Proceedings of an International Workshop, Bremen, Germany, March 4-8, 1991, edited by R. Duit, F. Goldberg, and H. Niedderer (IPN, Kiel Germany, 1992) 110-128;  
S. Vosniadou, "Capturing and modeling the process of conceptual change," *Learning & Instruction*, (4), 45-69, 1994;  
I. A. Halloun and D. Hestenes, "Common sense concepts about motion," *Am. J. Phys.* **53** (11), Nov. 1985.
- <sup>6</sup> The term mental model is used in a general sense to represent types of ideas, views, etc. held by students.
- <sup>7</sup> See reference 2.
- <sup>8</sup> See references 2 and 3.
- <sup>9</sup> See references 2 and 3.
- <sup>10</sup> See references 2 and 3.
- <sup>11</sup> See Appendix B in reference 3.
- <sup>12</sup> The plotted density is adjusted using  $10 \times \log(\mathcal{L} \times \rho + 1)$ , which removes the infinity of the logarithm by introducing a uniform offset equal to  $1/\mathcal{L}$ . When  $\mathcal{L}$  is large ( $5 \times 10^6$ ), this error is ignorable.
- <sup>13</sup> See reference 2.
- <sup>14</sup> Each class has 100 students, which is not a large set. The probability for an equal distribution is very small ( $\sim 10^{-6}$ ).
- <sup>15</sup> The peak's position can also be obtained analytically in terms of the variance of the random number generator; however, we found the numerical result is adequate for the purpose of this study.
- <sup>16</sup> Both eqs. (8) and (9) have an indefinite point at  $S = 100\%$  which makes  $C_{\text{Max}} = C_{\text{Min}}$ . In our calculation, we manually set the value of  $\Gamma$  and  $C_V$  to be 1 at  $S = 1$ .