

# A two-level system model for score-based measurement

Lei Bao

*Department of Physics, The Ohio State University,*

*174 W 18<sup>th</sup> Ave., Columbus, OH 43210*

## **Abstract**

The normalized gain (or the g-factor) has been widely used in physics education community as an assessment measure for student performance. In particular, it allows researchers to compare different instructions using classes with different initial states. Systematic differences were identified by R. Hake with thousands of students,<sup>1</sup> which show that classes, however different initially, tend to have similar values of g-factor when going through similar types of instruction, i.e., classes with traditional instructions often have systematically lower g's than classes with research-based instructions. These results indicate that g-factor can remove the individualities of the classes and reflect cleaner information about the instruction. However, the question of why the g-factor has this feature is still not well understood. In this paper, a physical model for the g-factor is proposed based on the context dependency of learning that may explain the physics behind the g-factor.

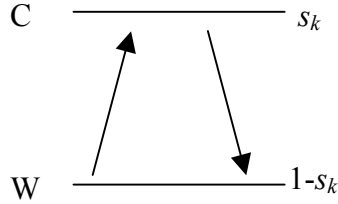
Three issues will be discussed in this paper. First, a model is developed to describe the score-based measurement of student knowledge and the change of the knowledge. This model leads to an explicit mathematical formulation of the possible dynamics of learning as reflected by the g-factor. Finally, other types of calculations are discussed and compared with the g-factor.

## **I. Student knowledge structure as reflected by score-based measurement**

In general, a student's knowledge on a particular concept topic can be considered with three categories: 1) scientifically correct knowledge, 2) scientifically incorrect knowledge, and 3) lack of knowledge. Depending on the nature of a particular measurement being conducted, the three components can result in different outcomes both independently and in combination. For example, if students' mental models are analyzed as part of the assessment (e.g. Model Analysis)<sup>2</sup>, the correct and incorrect knowledge is then reflected in the probability distribution for a student to use the different types of mental models (which can be correct, partially correct, incorrect, etc.). The lack of knowledge often results in a larger probability on a null model.<sup>3</sup> In the case of score-based measurement, the correct knowledge is measured with the successfulness of a student's giving correct answers. In this case, the 2<sup>nd</sup> and 3<sup>rd</sup> categories are collapsed together to represent the student's inability to provide correct answers, which is reflected by the complementary part of score (1-score). (A score used in this paper is always scaled to have a value between 0 and 1.) The measurement is then reduced to a single effective dimension. Apparently, many uncertainties are involved in such reduction of dimensions, but the benefit is the simplicity in application.

Now let's turn to the question of what can be represented by a measured score. Due to the complexity of cognitive process, it is very difficult to consistently attribute an actual mental construct to the score. The context features of test questions may activate different pieces of a student's knowledge as well as random human errors. Before we can parse out all the different possible mechanisms in a problem-solving process, these act as "hidden" variables and create uncertainties in the measured results. To get around this problem, let's consider the student's knowledge as reflected in the measurement space, i.e., we categorize the pieces of knowledge to be correct and/or incorrect only based on the measured results and don't attempt to probe why/how a particular result is produced. It is a very crude method but one can hardly go any further with score-based measurement. So in the score space, the knowledge will be identified with measured-correct, and measured-wrong (it may not be incorrect knowledge but is just not measured correct).

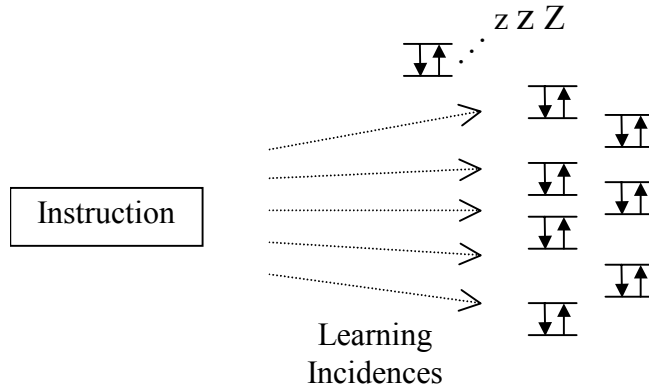
Then for a single test, one can compile a list of pieces of measured-correct knowledge that are embedded in the test. In this case, a student’s score can be interpreted as the quantity of the measured-correct knowledge that the student has. Since the total number of pieces of knowledge is fixed for a particular test, the score space is conserved. Thus the complementary part of the score ( $1-\text{score}$ ) gives the student’s measured-wrong knowledge.



Based on the above discussion, a single student’s measured knowledge in the score space can be represented with a “two-level” system (see Figure 1), where the C level (upper level) represents the measured-correct knowledge and its density is given by the score. The W level (lower level) represents the measured-wrong knowledge and its density is given by the complementary part of the score. A student’s state of measured-knowledge can be described with the densities on the two levels. Note that in a conserved score space, the two-level system only has one effective dimension.

**II. Interactions between instruction and students**

Then consider the whole instruction as being consisted of many learning incidences for a student. A learning incidence can be any activity introduced by the instruction (e.g., a part of the lecture, a lab, a homework problem, etc.). Each learning incidence has a probability to change a student’s state in some manner. Such changes can be viewed as transitions between the two levels of a student’s measured-knowledge. (Assume that we can perform such real time measurements). As a physics analogy, we can consider the individual students as identifiable particles (stuons) each with a unique two-level system. The triggers that cause the transitions in the two-level systems are a “stream” of learning incidences (lions) from the instruction (see the cartoon in Figure 2). Although the individual learning incidences have unique features in areas such as content of knowledge, learning activities, instruction format, etc.; in the model proposed here, we do not identify the individual learning incidences and treat them as an ensemble with a distribution of different features.



For a single student going through the stream of learning incidences, we can consider two types of transitions between the two levels of the student’s measured-knowledge state, an excitation process (increase of one’s measured-correct knowledge) and a decay process (decrease of one’s measured-correct knowledge). Apparently, for a class going through the same instruction, different students often interact differently with the instruction and show different changes on their measured-knowledge states. The differences of students’ interaction can be modeled with different excitation and decay “coefficients”.

### III. Change of a student's knowledge

The changes of a student's measured-knowledge state can be in a very complicated relation with the student's current state of knowledge and the characteristics of the instruction. Due to the limitations of score-based measurement, most of these issues cannot be explicitly identified and used in the modeling. Therefore, we can only consider the variables that can be obtained with scores, which actually gives only one variable – the score. Now we want to know how (or if) this variable is related to any of the possible processes in a student's learning.

It is well understood that students' existing knowledge can affect their learning in many different ways both constructively and as obstacles. Specifically, we can consider four processes with respect to the changes of a student's measured-knowledge.

1. E-I: An increase of measured-correct knowledge due to the help of the student's existing measured-correct knowledge. For example, a student may understand the correct knowledge to some level and can apply it to specific situations but may fail to apply in other contexts. In this case, knowing the correct knowledge at special situations can be helpful for the student to develop a more coherent understanding across different contexts.
2. E-II: An increase of measured-correct knowledge due to improvement of the student's measured-wrong knowledge. This reflects the processes where students restructure their incorrect initial knowledge and learn certain new knowledge that they have not encountered before.
3. D-I: A decrease of measured-correct knowledge due to the instability of some of student's existing measured-correct knowledge. For example, it is possible for the student to initially use scientifically incorrect knowledge to create a correct response and then gets confused during instruction and fails to provide a correct answer later.
4. D-II: A decrease of measured-correct knowledge due to the interactions from the student's existing measured-wrong knowledge. For example, it is not uncommon for students to develop fragmented understanding of a concept. A student may simultaneously possess two contradictory ideas about a single concept and treat them as separate issues (usually, each of the ideas is tied to a different set of contexts). Then during the instruction if the student realizes the contradiction and needs to reconcile the two ideas, it is possible that the incorrect idea may win. (If the correct one wins, it corresponds to the E-I process.)

With a score-based measurement, we won't be able to identify the specific pieces of knowledge that may contribute to the different changes. All the possible factors are collapsed into the score, which is considered to measure a wide variety of issues. (To reduce systematic biases, the measurement instrument needs to include multiple questions with diverse context settings so the approximation with scores can actually stand.) Then the changes of a student's measured-knowledge state are treated as to occur randomly during the instruction with a distribution of probabilities for the different processes to take place in each learning incidence.

For example, consider the process of E-II. A single student's measured-wrong knowledge is the result of many issues that cannot be identified by the score. These issues are often tied to specific contexts containing both content-based factors and learning environment factors. As the student going through instruction, which provides a sequence of learning incidences with diverse contexts (e.g., different content, learning formats, etc.), changes of the student's knowledge may occur in different ways at any learning incidences; but the actual processes of such changes as well as the specific learning incidences at which the changes take place remain largely uncertain to an external observer with a limited probing capability. It is also possible for a particular change of knowledge to be the result of an integrative effect of multiple learning incidences, in which case, the contributions of the individual incidences are even more difficult to determine. Then consider a large class of students with different background. The actual changes of these students' knowledge states during the instruction are nearly impossible to precisely determine with current technology. This leaves us little options but to treat the changes of students' knowledge states as an ensemble phenomenon without the details of the actual microscopic processes.

Now we can summarize two perspectives on the interaction between students and instruction: (1) the instruction provides a sequence of learning incidences with diverse context settings; and (2) the changes of individual students' knowledge states are the results of interactions with the learning incidences and can occur in any single incidence and or in an integration of multiple learning incidences.

#### IV. A mathematical function of the change of a student's knowledge

Still consider an E-II process. Suppose the measured-wrong knowledge reflects largely the student's scientifically wrong knowledge and the lack of appropriate knowledge. The probability of an improvement (decrease in amount) on the measured-wrong knowledge can be related primarily to two issues: the effectiveness of interactions with learning incidences, and the total quantity of the measured-wrong knowledge. In a constant steam of learning incidences, this probability reflects the rate of possible changes per learning incidence. Obviously, this probability is positively correlated to the effectiveness of the instruction on the particular student, and to the quantity of the measured-wrong knowledge that the student has – i.e., if a student has more problems (including incorrect knowledge and lack of knowledge) to improve, it is more likely for this student to encounter a learning incidence that improves one of his/her problems than a student who have a similar instruction effectiveness but less problems to be improved (has less new things to learn).

Suppose one can perform “identical” measurements during the instruction. The results of such measurements will produce a time dependent relation for the student's score  $s_k(t)$ . Denote  $\alpha_k$  to represent the effectiveness of the instruction on improving the  $k^{\text{th}}$  student's measured-wrong knowledge. The rate of possible changes in E-II processes per learning incidence can be represented with:

$$\text{E-II: } \frac{ds_k}{dt} = \alpha_k \cdot (1 - s_k)$$

Here,  $t$  represents the span of the series of learning incidences rather than the actual time. The rate of change of the measured-correct knowledge represents the ratio of the number of pieces of measured-correct knowledge changed per learning incidence over the total number of pieces of measured-correct knowledge as defined by the test. In theory,  $\alpha_k$  can have a very complicated relation with the student's existing knowledge and the instruction. Here we only use the simplest linear relation to describe the process. As a physics analogy, one can consider the quantity of measured-wrong knowledge as the student's “cross-section” for impacts of learning incidences, where  $\alpha_k$  gives the “excitation” coefficient for each impact (we may call it the *instruction impact coefficient*).

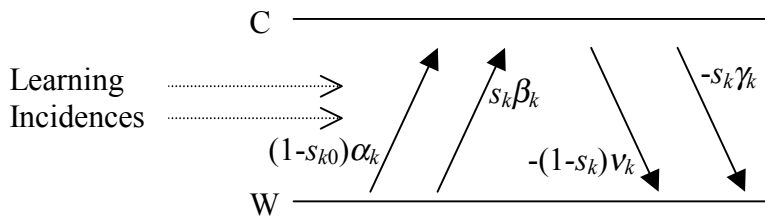
Similarly, for the other processes, we can write:

$$\text{E-I: } \frac{ds_k}{dt} = \beta_k \cdot s_k$$

$$\text{D-I: } \frac{ds_k}{dt} = -\gamma_k \cdot s_k$$

$$\text{D-II: } \frac{ds_k}{dt} = -\nu_k \cdot (1 - s_k)$$

Here, all coefficients are positive. The different processes are shown in Figure 3.



Combining all four processes together, we can write:

$$\frac{ds_k}{dt} = (1 - s_k) \cdot (\alpha_k - \nu_k) + s_k \cdot (\beta_k - \gamma_k) \quad (1)$$

Now let's think about if any of the processes may be dominant for a typical population of undergraduate college students. In physics education, large amount of research has documented that students' incorrect knowledge plays a significant role in their learning.<sup>4</sup> Less has been studied on how a students' correct knowledge may help them learn new concepts and how such knowledge may change inappropriately as a result of instruction. Based on literature, it is assumed that the effects of students' correct knowledge on learning new knowledge are less significant than the effects of incorrect knowledge. Further, D-II represents a process that the correct knowledge decreases due to the interactions with the incorrect knowledge. This is also less popular than the situation where it is just difficult to change a piece of incorrect knowledge (which is part of the E-II process). Here, it is assumed that the difference between the measured-knowledge of a student and the student's actual knowledge relevant to the measurement is ignorable. With the above approximations, the processes of E-I, D-I and D-II are ignored and Eq. (1) is reduces to

$$\frac{ds_k}{dt} = (1 - s_k) \cdot \alpha_k \quad (2)$$

The solution of Eq. (2) is found to be

$$s_k(t) = 1 - (1 - s_{k0})e^{-\alpha_k t} = 1 - z_k \cdot e^{-\alpha_k t} \quad (3)$$

where  $s_{k0}$  represents the student's initial score (pretest score) and  $z_k$  is the difference between  $s_{k0}$  and the perfect score (which is 1.0).

Suppose the instruction lasted for a duration of  $T$ . Then the initial and final scores of the student can be obtained as below.

$$\begin{aligned} s_k(0) &= 1 - z_k \\ s_k(T) &= 1 - z_k \cdot e^{-\alpha_k T} \end{aligned} \quad (4)$$

Notice that during instruction, a single student can acquire knowledge at any learning incidences that may be random in time domain. The relation described in Eq. (1) represents the ensemble results of large amount of interactions with a continuous stream of learning incidences and should not be regarded as the exact pattern that a particular student will follow during a short period of instruction.

For a single student, we can calculate the normalized gain:

$$g_k = \frac{s_k(T) - s_k(0)}{1 - s_k(0)} = \frac{z_k - z_k \cdot e^{-\alpha_k T}}{z_k} = 1 - e^{-\alpha_k T} \quad (5)$$

As we can see, by using the expression of the g-factor, one can remove the effects of the individual's initial score and retain only the change factor, which is closely related to instruction. Thus it provides a way to evaluate the effectiveness of different instructions with less interference from the variations of initial states of different populations.

Using Eq. (5), one can calculate  $\bar{g}$ , the class average of the individuals' gains:

$$\bar{g} = \frac{1}{N} \sum_{k=1}^N 1 - e^{-\alpha_k T} = 1 - \frac{1}{N} \sum_{k=1}^N e^{-\alpha_k T} \quad (6)$$

Now define  $\bar{\alpha}$  as population's instruction impact coefficient, where

$$e^{\bar{\alpha} T} = \frac{1}{N} \sum_{k=1}^N e^{-\alpha_k T} \quad (7)$$

We then have

$$\bar{\alpha}T = \ln\left(\frac{1}{1-\bar{g}}\right) \quad (8)$$

The measure of  $\bar{\alpha}T$  can provide an estimation on the integrative (over instruction) probability/rate of how a learning incidence may help a typical student increase his/her measured-correct knowledge. One can further parse out the effects of  $T$  under appropriately prepared situations; but may introduce additional uncertainties.

Another method to calculate the normalized gain uses the class average scores of pre and post tests. We can call this the population g-factor, denoted with  $\mathbf{g}$ . The class average scores of pre and post tests are denoted with  $X$  and  $Y$  respectively and can be obtained as below.

$$\begin{aligned} X &= \frac{1}{N} \sum_{k=1}^N (1 - z_k) = 1 - \frac{1}{N} \sum_{k=1}^N z_k \\ Y &= \frac{1}{N} \sum_{k=1}^N (1 - z_k \cdot e^{-\alpha_k T}) = 1 - \frac{1}{N} \sum_{k=1}^N z_k \cdot e^{-\alpha_k T} \end{aligned} \quad (9)$$

The population g-factor can be calculated with

$$\mathbf{g} = \frac{Y - X}{1 - X} = \frac{\sum_{k=1}^N z_k \cdot (1 - e^{-\alpha_k T})}{\sum_{k=1}^N z_k} \quad (10)$$

Denote  $Z_0$  as the class average of the individual students' complements of initial score:

$$Z_0 = \frac{1}{N} \sum_{k=1}^N z_k \quad (11)$$

Rewrite Eq. (10)

$$\mathbf{g} = \frac{Y - X}{1 - X} = \frac{1}{N} \sum_{k=1}^N \frac{z_k}{Z_0} \cdot (1 - e^{-\alpha_k T}) = 1 - \frac{1}{N} \sum_{k=1}^N \frac{z_k}{Z_0} \cdot e^{-\alpha_k T} \quad (12)$$

Here we can see that the effects of the individual students' initial scores still exist, and the instruction impact factor is used in a weighted summation to calculate the population gain: students with above average pre scores contribute more towards the population gain than students with below average pre-scores. Since the probability for a student to acquire knowledge is reflected in the impact coefficient ( $\alpha_k$ ), the weighting in Eq. (11) make  $\mathbf{g}$  dependent on the individuals' pre scores. Therefore, to evaluate the instruction, it may be better to use  $\bar{\mathbf{g}}$ , the class average of the individuals' gains.

In the above discussion, the measurement noise from random human errors is ignored. When considering such noise, it is assumed that random errors do not have particular time dependence and can occur at any time in all possible ways. This type of error exists in the measurement process and doesn't affect a student's learning. Therefore, the relations in Eq. (1) remain the same. If one can take multiple "identical" measurement during the instruction to monitor the changes of a student's score with the learning incidences, the effects of random errors in each measurement should eventually average out. If we take Eq. (1) as the integrative results over a period of instruction, then we shouldn't include a random error term in it. On the other hand, when we use the pre and post scores, since there are only two measured results, we should include a term to represent the possible uncertainties of the measurements due to random human errors. Thus, Eq. (4) can be rewritten as

$$\begin{aligned} s_k(0) &= 1 - z_k + e_{k0} \\ s_k(T) &= 1 - z_k \cdot e^{-\alpha_k T} + e_{kT} \end{aligned} \quad (13)$$

where  $e_{k0}$  and  $e_{kT}$  represent the possible random errors of the measured scores on pre and post tests. Then the normalized gain will be affected by the noise:

$$\mathbf{g}_k = \frac{s_k(T) - s_k(0)}{1 - s_k(0)} = \frac{z_k - z_k \cdot e^{-\alpha_k T} + e_{kT} - e_{k0}}{z_k - e_{k0}} \quad (14)$$

Assume that the amplitude of the random noise is small compare to  $z_k$ , (i.e., students have low pretest scores.) Use  $e_k$  to estimate the  $k^{\text{th}}$  students “average” error amplitude, then we can approximate  $\mathbf{g}_k$  with a first order Taylor expansion:

$$\mathbf{g}_k = (1 - e^{-\alpha_k T}) \cdot \left(1 - \frac{e_k}{z_k}\right) + 2 \frac{e_k}{z_k} \quad (15)$$

As we can see, the uncertainty has two parts, a constant part and a part that rescales the signal (i.e., the term of  $1 - e^{-\alpha_k T}$ ). Note that the constant part represents a maximum estimation. Due to the random nature of human errors, results in real measurement can be significantly smaller than this estimation. The two parts of uncertainties are dependent on the student’s performances on the pre and post tests, and therefore will make  $\mathbf{g}_k$  dependent on the population’s pretest scores. The uncertainty is in the order of  $e_k/z_k$ .

### V. Other methods on calculating relative score gains

As shown in Eq. (6), the way to calculate  $\bar{\mathbf{g}}$  removes the population differences and gives a cleaner measure of the effects of the instruction. However, it is also interesting to see the characteristics of other types of methods. In statistics, there are a number of methods to calculate “relative change scores”.<sup>5</sup> Table 1 summarizes a few functions that may be useful in physics education research. It can be shown that all the forms can be expressed as a function of  $Y/X$ .<sup>6</sup> (Here  $Y$  and  $X$  are used to represent pretest and posttest scores in general – could be class averages or individual scores.)

Relative Score Functions	Mapping
$\frac{Y - X}{X}$	$\frac{Y}{X} - 1$
$\frac{Y - X}{Y}$	$1 - \frac{Y}{X}$
$\frac{Y - X}{(Y + X)/2}$	$\frac{Y/X - 1}{(Y/X + 1)/2}$

The function of g-factor was not included in the two references (4 and 5), which gave quite extensive discussion on relative change scores. Further, the g-factor cannot be expressed in terms of  $Y/X$ . It appears that the g-factor is not commonly used in standard statistics.

In physics education, one of the goals to use relative changes instead of absolute changes is to reduce the effects of systematic differences of different populations so that the effects of the external interventions can be better evaluated. The pick of a particular method to calculate the relative change implies certain models being assumed for the data. Now let’s see what kind of model may be implied by using  $Y/X$  or its derivative forms.

In order to remove the effects of population with  $Y/X$ , we need to be able to write a student’s pre and post scores in the form shown below:

$$s_k(T) = s_k(0) \cdot f(T) \quad (16)$$

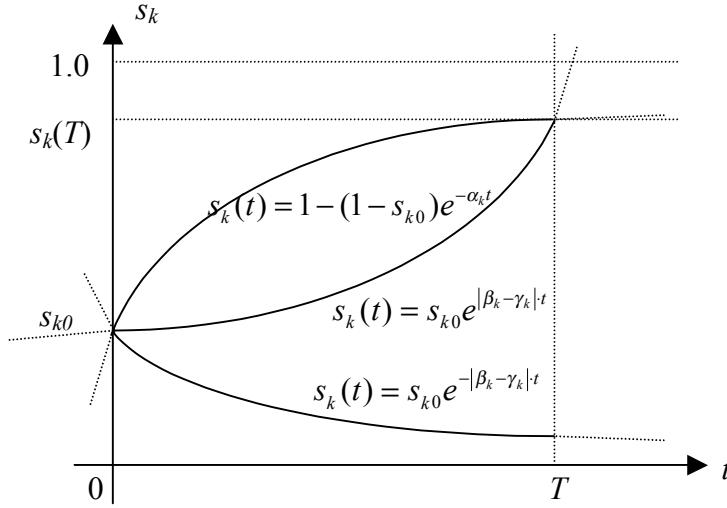
where  $f(t)$  is a time-dependent function. Consider the four approximations in Eq. (1). To have a solution in the form of Eq. (13), we need to remove the terms that are not in direct relation with  $s_k$ . Therefore, only the processes of E-I and D-I can remain, which gives

$$\frac{ds_k}{dt} = s_k \cdot (\beta_k - \gamma_k) \quad (17)$$

The solution to this equation is found to be

$$s_k(t) = s_{k0} e^{(\beta_k - \gamma_k)t} \quad (18)$$

As we can see, the model implied by  $Y/X$  suggests that the change of a student's knowledge is primarily related to the student's exiting measured-correct knowledge. Based on the research in education and physics education, this model is often less appropriate than the model behind the g-factor where students' incorrect knowledge is considered to the primary factor. In addition, the solution shown in Eq. (3) is convergent whereas the solution to Eq. (14) can be divergent (see Figure 5). Therefore, based on the underlying physical models and the mathematical descriptions, it is evident that using g-factor (in the form of  $\bar{g}$ ) can give a better evaluation of the effects of the instruction than other score-based methods discussed here.



## V. Discussions on technical issues and implications

### Basic assumptions for using g-factor

To consider g-factor as a measure for the effects of the instruction that is invariant to the population, one then assumes that the students' existing correct knowledge does not contribute significantly to the learning and the change of the students' knowledge is primarily due to the interaction of their incorrect and missing part of knowledge. Suppose the students' existing correct knowledge is also significant. Then Eq. (1) has to be in a more general form:

$$\frac{ds_k}{dt} = (1 - s_k) \cdot \alpha'_k + s_k \cdot \beta'_k = \alpha'_k - (\alpha'_k - \beta'_k) \cdot s_k \quad (19)$$

where  $\alpha'_k$  and  $\beta'_k$  represent the general dependence on  $(1-s_k)$  and  $s_k$  respectively. The solution to Eq. (16) is

$$s_k(t) = \frac{\alpha'_k}{\alpha'_k - \beta'_k} - \left( \frac{\alpha'_k}{\alpha'_k - \beta'_k} - s_{k0} \right) \cdot e^{-(\alpha'_k - \beta'_k)t} \quad (20)$$

Then  $g_k$  is found to be:

$$g_k = \frac{s_k(T) - s_k(0)}{1 - s_k(0)} = \frac{\left( \frac{\alpha'_k}{\alpha'_k - \beta'_k} - s_{k0} \right) (1 - e^{-(\alpha'_k - \beta'_k)T})}{1 - s_{k0}} \quad (21)$$

In this case, the effects of a student's pretest score ( $s_{k0}$ ) cannot be removed unless  $\beta'_k$  equals 0. However, if one can estimate the ratio between  $\alpha'_k$  and  $\beta'_k$  (i.e., between the effects of the incorrect knowledge on learning and the effects of the correct knowledge on learning), then Eq. (19) can be used to evaluate the instruction when both processes are considered.

$$h_k = \frac{s_k(T) - s_k(0)}{\frac{\alpha'_k}{\alpha'_k - \beta'_k} - s_k(0)} = \frac{(\frac{\alpha'_k}{\alpha'_k - \beta'_k} - s_{k0})(1 - e^{-(\alpha'_k - \beta'_k)T})}{\frac{\alpha'_k}{\alpha'_k - \beta'_k} - s_{k0}} = 1 - e^{-(\alpha'_k - \beta'_k)T} \quad (22)$$

As the other extreme, if one assumes that existing correct knowledge have the most significant effects, Eq. (22) reduces to one of the  $Y/X$  forms.

In this model, we used simple linear relations (see Eq. (1)). One can also assume other forms of relations with the score, which can result in different functions and the use of g-factor will also be affected.

### ***The implications on assessment and instruction***

Based on the data from over six thousands students, Hake was able to identify systematic differences between traditional instruction and research-based instruction. Based on the model proposed in this paper, Hake's results can be interpreted as that the research-based instruction provides learning incidences with higher probabilities to help students "fix" their incorrect knowledge. One can find that challenging and restructuring students' inappropriate conceptions are standard approaches in the research-based instruction such as Tutorials, Workshop Physics, etc.<sup>7</sup> The systematic differences in Hake's data show that the g-factor was able to separate the "instruction impact coefficients" with the students' "impact cross-sections", and therefore, provides an assessment on the effects of instruction that is relatively invariant across different populations. As we can see, the proposed model is consistent with the results of the experiment and the features of the instruction.

From research, we often observe that many students initially have low pretest scores on concept tests such as FCI, which provide large "target area" so it is not too difficult for instructors to have some impacts. As implied by this "collision" model, effective instruction not only needs to be able to make collisions with the target but also have to make more real ones – inelastic type preferred.

### **Acknowledgement**

The author wants to thank Professor E. F. Redish for many of his visionary suggestions and discussions. This work is supported in part by the NSF grant REC-0087788.

### **Endnotes and Reference**

- 
- <sup>1</sup> "Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," Hake, R. R., *American Journal of Physics*, **66** (1) pp. 64-74, 1998.
  - <sup>2</sup> Bao, L. (1999) "Dynamics of Student Modeling: A Theory, Algorithms, and Application to Quantum Mechanics," Ph.D. dissertation, University of Maryland.
  - <sup>3</sup> Bao, L. and Redish, E. F. (2001a) Concentration Analysis: A Quantitative Assessment of Student States, in press by PERS of AJP, July 2001;  
Bao, L. and Redish, E. F. (2001b) Model Analysis: Assessing the Dynamics of Student Learning, submitted to Cognition and Instruction.MA reference.
  - <sup>4</sup> McDermott, L. C. and Redish, E. F. (1999). Resource Letter PER-01: Physics Education Research. *Am. J. Phys.* **67**, 755-767.
  - <sup>5</sup> See p78 in *Analysis of Pretest-Posttest Designs*, Bonate, P. L., Chapman & Hall/CRC, Boca Raton, 2000.
  - <sup>6</sup> "How should relative change be measured?" L. Tornqvist, P. Vartia, and Y.Q. Vartia, *Am. Stat.*, **39**, 43, 1985.
  - <sup>7</sup> McDermott, L. C. and Shaffer, P. S. (1998). Tutorials in Introductory Physics, Prentice Hall, New York NY.