

Losing it: The Influence of Losses on Individuals' Normalized Gains

Kelly Miller^{1,4}, Nathaniel Lasry^{2,3,4}, Orad Reshef¹, Jason Dowd⁵, Ives Araujo^{4,6} & Eric Mazur^{4,5}

1 McGill University, Montreal Canada

2 Center for the Study of Learning and Performance, Montreal Canada

3 John Abbott College, Montreal Canada

4 School of Engineering & Applied Sciences, Harvard University, Cambridge MA, USA

5 Department of Physics, Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138

6 Physics Institute, UFRGS, Av. Bento Gonçalves 9500, Porto Alegre-RS, Brazil (CAPES - Proc. n^o: BEX 2271/09-5)

Abstract. Researchers and practitioners routinely use the normalized gain (Hake, 1998) to evaluate the effectiveness of instruction. Normalized gain (g) has been useful in distinguishing active engagement from traditional instruction. Recently, concerns were raised about normalized gain because it implicitly neglects retention (or, equivalently, "losses"). That is to say, g assumes no right answers become wrong after instruction. We analyze individual standardized gain (G) and loss (L) in data collected at Harvard University during the first five years that Peer Instruction was developed. We find that losses are non-zero, and that losses are larger among students with lower pre-test performances. These preliminary results warrant further research, particularly with different student populations, to establish whether the failure to address loss changes the conclusions drawn from g .

Keywords: Force Concept Inventory, normalized gain, gain, loss, Peer Instruction.

PACS: 01.30.Cc, 01.40.Fk, 01.40.gf

INTRODUCTION

Simple metrics of learning are succinct and readily consumable to the broader community of physics instructors, which makes them appealing to physics education researchers. Hake's gain – or normalized gain, as it was originally labeled – is a prime example of one such metric [1]. Introduced to physics education by Richard Hake in 1998, this metric is frequently used to evaluate student improvement when a test is administered at the beginning (pre-test) and at the end (post-test) of a course. Hake's gain, g , is defined as the ratio of the difference in total score to the maximal possible increase in score:

$$g = \frac{post - pre}{1 - pre}. \quad (1)$$

The terms *pre* and *post* represent the percent grade received on the test at the beginning and at the end of the course, respectively. When calculating the average

g for an entire class, the instructor has two choices: calculate g for each student and then average the values, or average *pre* and *post* for the class and then calculate g . The former approach is more rigorous, though a perfect pretest score results in division by zero. The latter approach is often simpler to calculate, though it masks important information about individual students. In this paper, we focus on individual student gains, and therefore rely solely on the former approach.

Despite its prolific use in evaluating conceptual change (as measured by multiple-choice surveys) across groups and amongst individuals, there are several limitations associated with using Hake's gain as an unbiased metric. In cases where the post-test score is not higher than the pre-test score, the normalized gain yields a number ranging from $-\infty$ to 0. In other words, when loss is normalized with respect to possible gain, the actual value for g does not lend itself to sensible interpretation. Furthermore, when calculating an average value of g , losses are potentially weighted much more heavily than gains. In an effort to

address this problem, Marx and Cummings suggest changing the denominator of g from $1-pre$ to pre when Hake's gain is negative, so that g is normalized with respect to possible loss [2].

The low pre-test score bias and asymmetric range of scores make interpretation of Hake's gain problematic [2]. However, even positive values of g can be unclear. Recently, David Dellwo has drawn attention to the fact that g does not incorporate retention of knowledge (and thereby losses) during the instructional period [3]. For example, if a student answers the first 40 questions of a pre-test correctly and the last 60 incorrectly, and then reverses this pattern on the post-test, the student's normalized gain of 0.33 does not illustrate that conceptual gains are accompanied by considerable losses.

Dellwo proposes two alternative metrics, instead of g , for measuring the conceptual change that occurs during the course of an instructional period. These metrics address both gain and retention. Dellwo defines gain (G) as a measure of the likelihood that a mistake on the pre-instruction test is corrected on the post-instruction test. Loss (L) is defined as the likelihood that a correct response on the pre-test is changed to an incorrect answer on the post-test.

In this paper, we examine the difference between the Hake gain, g , and Dellwo's metrics, G and L , in evaluating the effectiveness of physics instruction, as measured using the Force Concept Inventory. In an actual student population, if losses are negligible, the difference between g and G will be very small; however, we present data which suggests that the difference is not negligible. We carried out the following preliminary analysis using data collected at Harvard University from 1991 to 1996 in an introductory physics course where Peer Instruction was being implemented.

THEORETICAL FRAMEWORK

Hake introduced the normalized gain [1] so that the learning gains (as measured by multiple-choice pre- and post-tests) of students with different pre-test scores could be evaluated and compared in a meaningful way. Hake showed that instructional approaches could be distinguished on the basis of this metric: $\langle g \rangle$ tends to be lower in traditionally-taught courses and higher in active engagement courses. Yet, if losses are not taken into account, the use of $\langle g \rangle$ may yield misleading conclusions about the level of conceptual change that is taking place. What is the difference between g and Dellwo's G and L ?

Difference between g , G and L

To illustrate the role of losses as we describe these metrics, we take a slightly different approach from Dellwo. We define four possible transitions that can occur between the pre-test and the post-test. These transitions (presented in Table 1) consist of: right to right (RR), right to wrong (RW), wrong to wrong (WW) and wrong to right (WR).

Table 1: Four possible transitions of answers between the pre-test and post-test.

Pre-test	Post-test	Transition
Right	Right	RR
Wrong	Wrong	WW
Right	Wrong	RW
Wrong	Right	WR

Hake's gain can therefore be re-expressed as a function of these transitions:

$$g = \frac{WR - RW}{WW + WR}. \quad (2)$$

Indeed, $post - pre$ is equivalent to $WR - RW$, and $1 - pre$ is equivalent to $WW + WR$. Similarly, G can be expressed as

$$G = \frac{WR}{WW + WR}, \quad (3)$$

and L can be expressed as

$$L = \frac{RW}{RR + RW}. \quad (4)$$

G is normalized with respect to the potential gain, and L is normalized with respect to the potential loss. However, in g , both the gain and loss are normalized with respect to the potential gain.

Hake's gain can be expressed as a function of G and L as follows:

$$g = G - \gamma L, \quad (5)$$

where

$$\gamma = \frac{RR + RW}{WW + WR}. \quad (6)$$

RESULTS AND DISCUSSION

The relative number of transitions over the five years of data is shown in Figure 1. The percentage of answers changed from right to wrong, while small (3%), is not zero (as Hake's gain assumes).

FCI AVERAGE TRANSITIONS 1991-1996

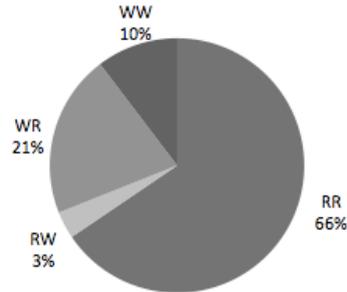


Figure 1: Percentage of total answers during the five year period for each of the four transitions (RR, RW, WW, WR) between pre-and post-test.

Dellwo opts to discuss the “renormalized” loss, γL , rather than L , so as to maintain a clear link to Hake's gain. However, this renormalized loss suffers from the same problems indicated by Marx and Cummings [2]: the scale is non-symmetric and does not lend itself to interpretation. In other words, Hake's gain simply does not sensibly account for loss, even when gain and loss are separated. If, however, we consider G and L , rather than γL , these problems are resolved. We investigate these metrics using actual student data. In instances where students' pre-test scores are 100% and 0%, calculations of G and L , respectively, result in division by zero; therefore, we assign G a value of 0 for students with scores of 100% and L a value of 0 for students with scores of 0% (we did not encounter the latter case).

Equation 2 above can be expressed as the difference between two ratios:

$$g = \frac{WR}{WW + WR} - \frac{RW}{WW + WR}. \quad (7)$$

Though the first term corresponds to G , the second term has no apparent meaning. This lack of meaning of the second term in Equation 7 reveals an implicit assumption: Hake's gain assumes the number of RW transitions is zero.

We address the following question: does measuring performance change using G and L provide a more complete picture than using g ?

METHOD

The Force Concept Inventory (FCI) [4] was administered at the beginning and at the end of each semester during the first five years that Peer Instruction was being developed at Harvard University [5]. Between 1991 and 1994, the original 29-item version of the FCI was used [4]. Between 1995 and 1996, the 30-item version of the FCI was used. Between 1991 and 1996, 894 students completed both the pre-test and post-test of the FCI. The student population size per year ranged from $n=158$ to $n=215$.

For each year of data, we recorded the number of transitions (RR, RW, WW, WR) for each student. From this, we calculated g , G and L for each student and computed yearly averages.

Table 2 shows that both g and G correspond to improvement in students' performance. Both metrics follow the same trend, as L is relatively constant over these years. If L were negligibly small, G would be approximately equal to g .

Table 2: Hake's Gain and $G-L$ across the first five years of peer instruction at Harvard University.

YEAR	Gain (G)	Loss (L)	g
1991	0.66	0.07	0.45
1993	0.69	0.06	0.52
1994	0.73	0.06	0.59
1995	0.82	0.07	0.71
1996	0.79	0.06	0.67
AVG	0.74	0.06	0.59

Figure 2 illustrates the relationship between each student's G and g . This plot shows that the RW transitions, or losses, lead to lower values of g over the full range of values of G .

If L were negligible, all of these points would appear along a straight line, $G = g$. The scatter away from that line indicates that non-zero losses occur over the entire range of possible gains. Of the students included here, 382 had no losses ($L = 0$) and 512 had some losses ($L > 0$). As indicated in equations 5 and 6, the extent to which losses affect the Hake's gain of an individual depends on the pre-test performance; if a student answers only a few questions wrong on the pre-test, losses will have a much larger effect than for a student who answers only a few questions right. Thus,

initially-high-performing populations of students are more susceptible to diminished Hake's gains than initially-low-performing populations.

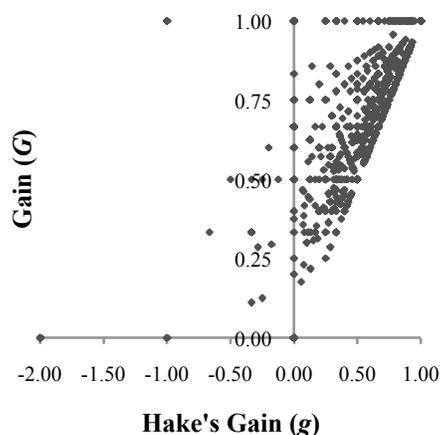


Figure 2: We have plotted G vs. g to demonstrate that RW transitions lead to lower values of g over the full range of values of G .

To further investigate the relationship between Hake's gain and losses, we present the average normalized loss for each quartile of the population, according to pre-test score, in Table 3.

Table 3: Average normalized loss for each quartile of the population according to pre-test score

Quartile	Pre-Test (%)	Average L
1	0.20-0.57	0.11
2	0.57-0.70	0.06
3	0.70-0.83	0.05
4	0.83-1.00	0.02

Students with lower pre-test scores (in the first quartile) have higher losses than students with higher pre-test scores. This suggests that, within this population of Harvard University undergraduates, losses vary with pre-test score. This data also suggests that some of the observed values of g may result from a combination of smaller gains and moderate losses, as well as larger gains and small losses with a larger impact because of high pre-test scores.

CONCLUSION

Concerns have recently been raised about the normalized (Hake's) gain because it implicitly assumes that losses are zero. We compared Hake's gain to alternative metrics, G and L , that account for retention/losses in addition to performance gains. Using FCI data from Harvard University over the first

five years of instruction, we have shown that losses are, indeed, non-zero and not necessarily negligible. In particular, students who score relatively poorly on the pre-test show larger losses in performance. We have demonstrated that the implicit assumption of Hake's gain is false.

However, upon making this claim, two questions immediately arise. First, do these "losses" represent actual conceptual losses, or do they result from correct guesses on the pre-test that, by chance, became incorrect on the post-test? Second, do these losses change the conclusions that have been drawn using Hake's gain, specifically regarding the effect of Peer Instruction on FCI performance? These questions will be addressed in further analysis of data collected from multiple student populations.

REFERENCES

1. Hake, R. R. (1998). "Interactive-Engagement vs. Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses." *Am. J. Phys.* **66**(1): 64-74.
2. Marx, J.D., Cummings, K., Normalized Change. *Am. J. Phys.*, 2007. **75**(1): p. 87-91.
3. Dellwo, D. (2009). Reassessing Hake's Gain. Preprint, available on request.
4. Hestenes, H., Wells, M., Swackhamer, G. (1992). "Force Concept Inventory." *The Physicist Teacher* **30**: 141-158.
5. Mazur, E. (1997). Peer Instruction: A User's Manual, Prentice Hall.