Journal of the Scholarship of Teaching and Learning, Vol. 10, No. 1, January 2010, pp. 55 – 67.

Course assessment using multi-stage pre/post testing and the components of normalized change

David R. Dellwo[1]

Abstract: A multi-stage pre/post testing scheme is developed to gauge course effectiveness using gain and loss components of normalized change. The components, unlike normalized change itself, can be used to distinguish courses that promote acquisition as well as retention of information from courses that promote acquisition at the expense of retention or retention at the expense of acquisition. The technique is employed to study the effectiveness of a course in differential calculus taught using the studio method, a form of interactive engagement.
Keywords: course assessment, multi-stage, pre/post testing, normalized change, normalized gain, normalized loss, studio method, interactive engagement.

I. Introduction.

Assessment of learning is a recurrent and sometimes controversial theme in higher education. The literature is replete with conflicting advice on how best to conduct an assessment, see Hake (2004, 2006) and Suskie (2004a, 2004b). However, pre- and post-testing evaluation is often cited as a commonsense approach. Perhaps no one has expressed this point of view more graphically than Bond (2005), who wrote in a Carnegie Perspective:

> *If one wished to know what knowledge or skill Johnny has acquired over the course of a semester, it would seem a straightforward matter to assess what Johnny knew at the beginning of the semester and reassess him with the same or equivalent instrument at the end of the semester.*

Theoretical justification for the technique was provided by Willet (1989a, 1989b, 1994, 1997) and Rogosa (1995). In particular, they demonstrated that additional rounds of pre- and post-testing dramatically improve the method's reliability.

The multi-stage assessment scheme employed here partitions the course into several instructional periods. Each period is bracketed by pre-instruction and post-instruction tests, with the post-test for one period serving as the pre-test for the next period. This arrangement creates an opportunity to study the marginal (snapshot) effectiveness of individual instructional periods or alternatively to combine individual periods and study the cumulative (longitudinal) effectiveness of several combined instructional periods.

The two analyses provide information on different aspects of course effectiveness. A cumulative analysis is used to determine whether repeated exposure to course material over multi-periods of instruction increases the likelihood of students acquiring and retaining baseline knowledge. A marginal analysis is used to determine whether course design is flexible enough to

[1] Department of Mathematics and Science, United States Merchant Marine Academy, Steamboat Road, Kings Point, NY 11024, dellwod@usmma.edu

continually rebalance acquisition and retention efforts as student performance changes from one instructional period to the next.

The method used to quantify changes in performance is a definitive feature of any pre/post testing design. The following index is frequently used to measure the change in group performance from a pre-instruction to a post-instruction test.

$$g = \frac{\left\{\begin{matrix} average\ grade\ on\ the \\ post\text{-}instruction\ test \end{matrix}\right\} - \left\{\begin{matrix} average\ grade\ on\ the \\ pre\text{-}instruction\ test \end{matrix}\right\}}{100 - \left\{\begin{matrix} average\ grade\ on\ the \\ pre\text{-}instruction\ test \end{matrix}\right\}} \tag{1}$$

The ratio in (1), often referred to as normalized change, expresses the difference between average test scores as a fraction of the maximum possible difference between these scores.

Hovland et al. (1949) used (1) to quantify the effectiveness of instructional films. Hake (1998) used (1) to gauge the relative effectiveness of various instructional techniques employed in introductory physics courses. Cummings et al. (1999) used (1) to evaluate innovations in studio physics. Meltzer (2002) used (1) to explore the relationship between mathematics preparation and concept learning in physics. These important studies relied on the intuitive notion that when comparing two courses:

The course with the larger value of normalized change (g) is the more effective course. (2)

Unfortunately, as demonstrated here, this classic assessment rule can lead to counterintuitive conclusions.

This paper employs an alternate assessment rule obtained by decomposing normalized change (1) into component measures:

$$g = G - \gamma L \tag{3}$$

Here G is a normalized gain measuring the likelihood that a mistake on the group's pre-instruction test is corrected on the post-instruction test. Similarly, L is a normalized loss measuring likelihood that a correct response on the group's pre-instruction test is rendered incorrect on the post-instruction test. The non-negative parameter γ is a renormalization factor dependent on the population's pre-instruction performance. Consequently, (3) expresses normalized change (1) as the difference between two non-negative indices, normalized gain and renormalized loss. The decomposition (3) gives rise to an alternative assessment rule that avoids the counterintuitive conclusions associated with (2), and reads in part:

The course with the larger value of normalized gain (G) and smaller value of renormalized loss (γL) is the more effective course. (4)

The derivation of (3) is discussed in the next section. Section III discusses assessment standards and value-added measurement of effectiveness expressed in terms of the components of normalized change. Multi-stage assessment is discussed in section IV. The application is presented in section V. The concluding remarks in section VI discuss the implications of (3) for

past and future research.

II. Normalized Change and Its Components

Normalized change (1) for a group of N students taking a diagnostic test with M questions can be expressed in the following form:

$$g = \frac{\theta_{post} - \theta_{pre}}{1 - \theta_{pre}} \tag{5.a}$$

where

$$\theta_{pre} = \frac{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ correctly\ \ on\ the\ pre\text{-}instruction\ test \end{array}\right\}}{NM} \tag{5.b}$$

$$\theta_{post} = \frac{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ correctly\ \ on\ the\ post\text{-}instruction\ test \end{array}\right\}}{NM}. \tag{5.c}$$

The derivation of (3) is based on the following observation.

$$\left\{\begin{array}{l} Number\ of\ questions \\ students\ answer\ correctly \\ on\ the\ post\text{-}instruction\ test \end{array}\right\} - \left\{\begin{array}{l} Number\ of\ questions \\ students\ answer\ correctly \\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}$$

$$= \left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ correctly\ on\ the\ post\text{-}instruction\ test \\ and\ incorrectly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\} - \left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ incorrectly\ on\ the\ post\text{-}instruction\ test \\ and\ correctly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}$$

This observation together with definitions (5.b) and (5.c) imply

$$\theta_{post} - \theta_{pre} = G(1 - \theta_{pre}) - L\,\theta_{pre} \tag{6}$$

where

$$G = \frac{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer\ correctly\ on\ the\ post\text{-} \\ instruction\ test\ and\ incorrectly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}}{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ incorrectly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}} \tag{7.a}$$

$$L = \frac{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer\ incorrectly\ on\ the\ post\text{-} \\ instruction\ test\ and\ correctly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}}{\left\{\begin{array}{l} Number\ of\ questions\ students\ answer \\ correctly\ on\ the\ pre\text{-}instruction\ test \end{array}\right\}}. \tag{7.b}$$

The numerator in (7.a) is the number of questions on which students demonstrate a gain in knowledge and the denominator is the maximum possible gain. Consequently, the ratio G is a

normalized gain measuring the conditional probability (Ross, 2004) that a mistake on the group's pre-instruction test is corrected on the post-instruction test.

Similarly, the numerator in (7.b) is the number of questions on which students demonstrate a loss in knowledge and the denominator is the maximum possible loss. Consequently, the ratio L is a normalized loss measuring the conditional probability that a correct response on the group's pre-instruction test is rendered incorrect on the post-instruction test.

In summary, equation (6) expresses change in test score as a difference between the fraction of questions on which students demonstrate a gain in knowledge and the fraction on which they demonstrate a loss of knowledge. Finally, to obtain (3) define

$$\gamma = \frac{\theta_{pre}}{1 - \theta_{pre}} \tag{7.c}$$

and divide (6) by $(1-\theta_{pre})$. The scaling factor (7.c) is a non-negative parameter whose value is larger than 1 if $\theta_{pre} > ½$, equal to 1 if $\theta_{pre} = ½$, and smaller than 1 if $\theta_{pre} < ½$. The scale γ is referred to as the group's aspect ratio and specifies the odds that the group gives a correct answer on the pre-instruction test.

III. Value-Added Measurement of Course Effectiveness.

The following criteria are used in this study to assess the relative effectiveness of two courses (A and B).

i. *A is more effective than B if:*
$$\begin{cases} G_A > G_B \text{ and } \gamma_A L_A \le \gamma_B L_B \\ \text{or} \\ G_A \ge G_B \text{ and } \gamma_A L_A < \gamma_B L_B \end{cases} \tag{8.a}$$

ii. *A and B are equally effective if:* $G_A = G_B$ *and* $\gamma_A L_A = \gamma_B L_B$ (8.b)

iii. *A and B are not comparable if:*
$$\begin{cases} G_A > G_B \text{ and } \gamma_A L_A > \gamma_B L_B \\ \text{or} \\ G_A < G_B \text{ and } \gamma_A L_A < \gamma_B L_B \end{cases} \tag{8.c}$$

Notice, (8.a) restates (4) in algebraic form and defines a consistent ordering of courses in the sense that if A is more effective than B and B is more effective than C, then A is more effective than C. Also, (8.c) offers an assessment option not offered by (2): namely, some courses are not comparable.

If A is a more effective course than B in the sense of (8.a), then $G_A - G_B$ is a value-added measure of improved effectiveness due to larger gains, see (Suskie, 2004a). Also, $\gamma_B L_B - \gamma_A L_A$ is a value-added measure of improved effectiveness due to smaller renormalized losses experienced by students in the more effective course. Consequently,

$$g_A - g_B = (G_A - G_B) + (\gamma_B L_B - \gamma_A L_A) \tag{9}$$

is a value-added measure of the total improvement in effectiveness when (8.a) or equivalently (4)

applies and one course can claim the larger gains as well as the smaller renormalized losses.

On the other hand, (9) is not a measure of total improvement in effectiveness when (8.c) applies and neither course can claim both larger gains and smaller renormalized losses. In this case, one of G_A - G_B and $\gamma_B L_B$ - $\gamma_A L_A$ is positive while the other is negative; so (9) is the difference between two value-added measures:

$$g_A - g_B = (G_A - G_B) + (\gamma_B L_B - \gamma_A L_A) = \begin{cases} -(G_B - G_A) + (\gamma_B L_B - \gamma_A L_A) & \text{if } (G_A - G_B) < 0 \\ (G_A - G_B) - (\gamma_A L_A - \gamma_B L_B) & \text{if } (\gamma_B L_B - \gamma_A L_A) < 0 \end{cases} \quad (10)$$

That is, g_A - g_B is a difference between added effectiveness due to larger gains in one course and added effectiveness due to smaller renormalized losses in the other course.

Finally, in view of (10), the classic assessment rule (2) declares A more effective than B when either of the following applies.

- The added effectiveness due to smaller renormalized losses in A offsets the added effectiveness due to larger gains in B.
- The added effectiveness due to larger gains in A offsets the added effectiveness due to smaller renormalized losses in B.

Of course, neither of these alternatives can form the basis for a pedagogically sound strategy to improve learning.

IV. Multi-Stage Assessment.

Most pre/post assessment regimes employ a single instructional period bracketed by identical or nearly identical pre- and post-instruction tests. See (Hake, 1998), (Cummings et al., 1999), (Meltzer, 2002), (Libarkin et al., 2005), and (McConnell et al., 2006). Unfortunately, these single-stage methods, relying on two tests, cannot gather enough data to detect inevitable fluctuations in learning that result from imperfect acquisition and retention of course material. For example, a round of pre/post testing cannot detect a difference in performance between a student who never learns a key skill and a student who learns and then forgets that skill during the term. Similarly, a round of testing cannot distinguish between a student who retains pre-instruction knowledge throughout the term and a student who forgets and then relearns that knowledge during the term.

Multi-stage regimes track fluctuations in learning and refine the assessment process by combining several single-stage regimens. For example, the two-stage scheme diagramed in Figure 1 can detect a one-time loss and reacquisition of course material as well as a one-time acquisition and subsequent loss of material. It is important to note that the inter-period diagnostic test (T_1) serves as a post-instruction test for the first stage as well as a pre-instruction test for the second stage.

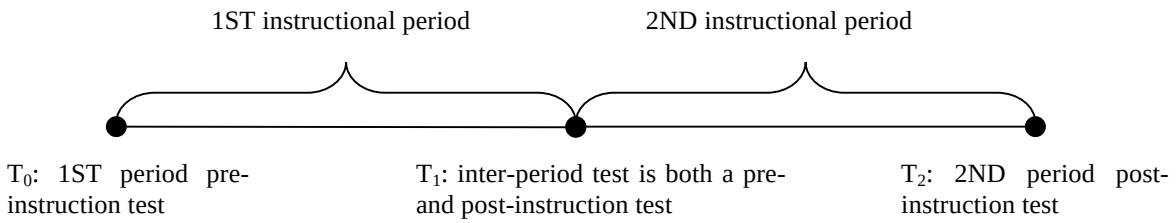

Figure 1. The first stage of a two-stage assessment scheme is bracketed by pre- and post-instruction tests T_0 and T_1. The second stage is bracketed by T_1 and T_2. The diagnostic tests are identical or nearly identical instruments designed to assess learning of key skills and concepts.

A. Marginal Analysis of Multi-Stage Schemes.

A marginal analysis uses the components of normalized change (3) to tabulate changes in performance relative to pre-instruction levels at each stage of a multi-stage scheme. This technique can be used to study variations in effectiveness from one instructional period to the next for a particular course. Or alternatively, the approach can be used to compare effectiveness of two courses during a particular instructional period.

Notice that for a marginal analysis the standard by which effectiveness is determined changes from period to period. For example, a marginal analysis of the two-stage scheme shown in Figure 1 might use gains and losses from T_0 to T_1 as well as from T_1 to T_2 to study variations in effectiveness for a single course. In this situation, improving effectiveness from the first to the second instructional period means the course was more effective in boosting learning relative to performance on T_1 than it was relative to performance on T_0.

As a second example, the marginal analysis of a two-stage scheme might be used to compare the effectiveness of two courses in promoting learning relative to T_0 as well as to T_1. In this situation, it may happen that one of the two courses is more effective in promoting learning relative to T_0 while the other is more effective in promoting learning relative to T_1.

B. Cumulative Analysis of Multi-Stage Schemes.

A cumulative analysis tabulates changes in performance over several successive stages of a multi-stage scheme by measuring gains and losses from the initial pre-instruction test to each of the subsequent post-instruction tests. In contrast to the marginal analysis, a cumulative analysis uses performance on T_0 as a fixed standard from which to gauge change over successive periods from T_0 to T_1, from T_0 to T_2, from T_0 to T_3, etc.

This technique can be used to compare a particular course's effectiveness during the single period from T_0 to T_1 with its effectiveness during the two periods from T_0 to T_2 in helping students rectify weaknesses revealed by their performance on the initial diagnostic test, T_0. Alternatively, the approach can be used to study the relative effectiveness of two courses in promoting learning over the first two, three, or more instructional periods following the initial diagnostic test.

V. Application.

The illustration presented here uses multi-stage pre/post testing to study the effectiveness of a course in differential calculus taught by the author to 125 plebes at the United States Merchant

Marine Academy over a seven-year period from the fall of 1999 to the fall of 2005. The course was taught using the studio method, a form of interactive engagement; see (Ecker, 1996a, 1996b), (Hake, 1998) and (Sokoloff et al., 1997). Methodology and other details are discussed first, in the next section, before presenting results.

A. Study Procedures.

This section discusses details of the teaching methods, the pre-instruction test, the pre- and post-instruction testing, as well as the student participants.

Teaching Method. Studio sections of the differential calculus course were taught using a modified form of the integrated lecture-laboratory-recitation design envisioned by Ecker (1996a, 1996b). In the classic studio setting, small groups of students work on in-class activities while receiving constant guidance and help from the instructor. There is essentially no lecture, and although there is homework, there is little use of out-of-class projects requiring written and/or oral reports.

The modified studio format used here incorporated instructor demonstrations of interactive Maple applications as well as out-of-class group projects. Instructor demonstrations exploited the computer algebra system's facility to perform "what if" scenarios in real time, giving students the opportunity to rapidly and efficiently test preconceptions and correct misconceptions. Often the demonstrations were used in conjunction with classic studio activities from the text *Studio Calculus* (Ecker, 1996b). On the premise that *teaching is the best way to learn*, the out-of-class group projects required studio students to construct interactive multimedia learning aids for use by other students.

Pre-Instruction Test. The pre-instruction (diagnostic) instrument included twenty-four multiple-choice questions concerning core concepts and skills associated with differential calculus. Specific topics include functions, limits, continuity, differentiation, as well as applications; see (Dellwo, 2000) for details. The diagnostic questions were typical of practice problems used to prepare for standardized exams. However, the questions were not vetted to the same degree as those developed for the Force Concept Inventory (Hestenes et al., 1992) or the more recent Calculus Concept Inventory (Epstein, 2007).

Pre- and Post-Instruction Testing. Each year, the pre-instruction test was administered on the first day of class. Post-instruction testing employed a regimen of quizzes intended to give each midshipman two post-instruction opportunities to answer the twenty-four diagnostic questions. Typically:

- The first post-instruction quiz was composed of diagnostic questions on topics covered in class during the first or second week of the term.
- The second post-instruction quiz was composed of questions repeated from the first post-instruction quiz and diagnostic questions on topics covered during the second or third week of the term.
- The third post-instruction quiz was composed of questions repeated from the first post-instruction quiz but not used on the second post-instruction quiz, questions repeated from the second post-instruction quiz, and diagnostic questions on topics covered in class during the third or fourth week of the term.

This process of combining and recombining questions from the pre-instruction test

continued until the end of the term and generally resulted in eight to ten quizzes of approximately seven questions each. Thus, none of the quizzes contained all the diagnostic questions given on the first day of class. Rather, each quiz contained a subset of diagnostic questions on topics discussed in class prior to giving that quiz. Consequently, the pre-instruction test (T_0) and the post-instruction tests (T_1, T_2) contained the same twenty-four diagnostic questions, but administered the questions in different ways. The pre-instruction test administered all the questions on the first day of class while the post-instruction tests administered the questions a few at a time on quizzes spread through out the term.

The pre-instruction test and the post-instruction quizzes were scored by assigning a numerical value of 1 to correct answers and a numerical value of 0 to incorrect answers. By term's end each student had accrued three scores for each of the twenty-four diagnostic questions. For a particular student answering a particular question these scores are:

- $S_0 = 1$ or 0 depending on whether the student answered the question correctly on the pre-instruction test.
- $S_1 = 1$ or 0 depending on whether the student answered the question correctly on the first post-instruction opportunity.
- $S_2 = 1$ or 0 depending on whether the student answered the question correctly on the second post-instruction opportunity.

There are NxM values of S_0 for a group of N students answering M questions. These values of S_0 determine the numerator in equation (5.b) and consequently the average grade on T_0. For example, the numerator for all studio sections under study was computed by summing the values of S_0 over all diagnostic questions and all studio students. Values of S_1 determine the numerator in (5.c) for the first post-instruction test and consequently the average grade on T_1. Similarly, values of S_2 determine the average grade on T_2.

The gain and loss components in equations (7.a) and (7.b) were computed in a similar fashion. For instance, the numerator in (7.a) for the studio gain from T_0 to T_2 was obtained by summing the values of $\max(S_2 - S_0, 0)$ over all diagnostic questions and all midshipmen in the studio sections.

Questions appearing on the calculus diagnostic test were never modified, but the optional choices were rearranged from time to time. Although this method has the disadvantage of using the same questions several times, it has the overriding advantage of eliminating any possibility that test questions revised for use on a later test could introduce ambiguities resulting in false gains and/or false losses. The technique eliminates the difficult, if not impossible, task of establishing equivalencies between seemingly similar questions.

Students. Midshipmen taking the studio course had some calculus in high school and demonstrated strong algebraic skills on a screening test given to all plebes. The course was taught in an electronic classroom that limited the number of students to twenty, but enrollment varied between fifteen and twenty students per year.

B. Effectiveness of Studio Calculus.

This section illustrates the use of a two-stage assessment scheme to study intra-term variations in effectiveness for a studio course in differential calculus. A marginal analysis is presented first, then a cumulative analysis.

Marginal Effectiveness. The results of a marginal analysis of gains and losses for the studio course are tabulated in Table 1. When reviewing the table, keep in mind that the inter-period test (T_1) is used as a pre- and a post-instruction test. Consequently, the aspect ratio changes from one period to the next. For example, the value 1.46 of the aspect ratio for the first instructional period is obtained from (7.c) using the average score on T_0. The value 3.09 for the second period is obtained from (7.c) using the average grade on T_1.

In addition, normalized gain and loss are conditioned on events that change from period to period. For example, in Table 1 the value 0.14 for the normalized loss from T_0 to T_1 means that 14% of diagnostic questions answered correctly on T_0 were answered incorrectly on T_1. The value 0.71 for normalized gain from T_1 to T_2 means that 71% of questions answered incorrectly on T_1 were answered correctly on T_2.

Table 1. Marginal analysis of gains and losses for the studio course. The initial pre-instruction test is designated by T_0, the first post-instruction test by T_1, and the second post-instruction test by T_2. Error estimates employ the standard deviant. The estimates for γ, γL, and g are based on conventional linearization techniques; see (Taylor, 1982) and the discussion in (Hake, 1998, p. 73).

Instructional Period	Aspect Ratio: γ	Normalized Loss: L	Renormalized Loss: γL	Normalized Gain: G	Normalized Change: g
T_0 to T_1	1.46±0.05	0.14±0.01	0.20±0.01	0.60±0.01	0.40±0.02
T_1 to T_2	3.09±0.13	0.07±0.01	0.22±0.02	0.71±0.02	0.49±0.03

Data in Table 1 indicates that renormalized loss was nearly constant from one instructional period to the next, with $\gamma L \approx 0.21$. Although nominal values of γL increased slightly, the increase is not large enough to be statistically significant. On the other hand, the difference in normalized gain is large enough to conclude that G increased from one period to the next.

In summary, for the studio course renormalized loss remained stable while normalized gain increased; and (8.a) leads to the conclusion that marginal effectiveness of the course improved from one period to the next. Moreover, according to (9), successive differences in nominal values of normalized change listed in Table 1 quantify the added effectiveness. The data indicates the studio course was 22.5% more effective in boosting learning relative to performance on T_1, when the odds of a correct answer were $\gamma_1 \approx 3$ and the average grade was 75%, than it was relative to performance on T_0, when the odds of a correct answer were $\gamma_0 \approx 1.5$ and the average grade was 60%.

Cumulative Effectiveness. Table 2 tabulates the results of a cumulative analysis of gains and losses for the studio course. When reviewing the table, keep in mind that for cumulative periods of instruction, change is measured relative to the initial diagnostic test, T_0. Consequently, the aspect ratio (7.c) has a fixed value.

Similarly, normalized gain and normalized loss are defined relative to performance on T_0. For example, the value 0.08 for the normalized loss from T_0 to T_2 means that 8% of diagnostic questions answered correctly on T_0 were answered incorrectly on T_2. The value 0.81 for the normalized gain from T_0 to T_2 means that 81% of diagnostic questions answered incorrectly on T_0 were answered correctly on T_2.

Table 2. Cumulative analysis of gains and losses for the studio course. The initial pre-instruction test is designated by T_0, the first post-instruction test by T_1, and the second post-instruction test by T_2. Error estimates employ the standard deviant. The estimates for γ, γL, and g are based on conventional linearization techniques; see (Taylor, 1982) and the discussion in (Hake, 1998, p. 73).

Instructional Period	Aspect Ratio: γ	Normalized Loss: L	Renormalized Loss: γL	Normalized Gain: G	Normalized Change: g
T_0 to T_1	1.46±0.05	0.14±0.01	0.20±0.01	0.60±0.01	0.40±0.01
T_0 to T_2	1.46±0.05	0.08±0.01	0.11±0.01	0.81±0.01	0.69±0.01

Inspection of Table 2 reveals that normalized gain was larger during the period from T_0 to T_2 than during the period from T_0 to T_1. Also renormalized loss was smaller from T_0 to T_2 than from T_0 to T_1. Consequently, the studio course was more effective in promoting learning relative to T_0 during the two instructional periods from T_0 to T_2 than during the single period from T_0 to T_1 by an amount equal to the difference in normalized change $\Delta g = 0.69\text{-}0.40 = 0.29$, see (9).

VI. Concluding Remarks: Was Hake Correct?

In 1998 Richard Hake published the results of a large survey of pre/post test data for introductory physics courses. He estimated the average normalized change for traditional (T) courses, those that made little use of interactive engagement (IE), at $g_T \approx 0.23$. He estimated the average normalized change for courses that made substantial use of IE methods at $g_{IE} \approx 0.48$. These findings are noteworthy because the estimate $g_{IE} \approx 0.48$ for interactive courses is almost two standard deviations above the estimate $g_T \approx 0.23$ for traditional courses. Hake (1998) concluded:

> *Classroom use of IE strategies can increase mechanics-course effectiveness well beyond that obtained in traditional practice.*

Hake's conclusion is certainly valid on the basis of (2), but is it valid on the basis of (8)? At present a complete answer cannot be given, since the average values of G and γL for traditional and interactive physics courses are not known. However, the decomposition (3) can be used to obtain a partial answer.

Since (3) is an identity, Hake's findings imply that assessment states for traditional courses must be distributed near the contour $g \approx 0.23$ in the $(G, \gamma L)$ plane. Furthermore, the mean assessment state for traditional courses must lie on this contour. Similarly, assessment states for interactive courses must be distributed near the contour $g \approx 0.48$ and the mean IE state must lie on that contour. See Figure 2.

If the mean IE state falls along the middle portion of the contour $g \approx 0.48$, shown in Figure 2, then (8) implies IE methods are more effective than traditional methods because on average they exhibit higher normalized gains and smaller renormalized losses. On the other hand, if the mean IE state falls along the upper or lower portions of the contour, as indicated in Figure 2, then (8) implies the two methods are not comparable because one method produces larger gains while the other produces smaller renormalized losses.

Thus, if (8), rather than (2), is used to gauge effectiveness, Hake's data implies that traditional physics courses cannot, on average, be more effective than interactive courses. That is, the traditional approach is either less effective than the interactive approach or the two methods are not comparable. Although this statement is not as strong as Hake's original statement, future efforts to determine the average values of G and γL for traditional and

interactive physics courses may make it possible to say more, even more than Hake originally envisioned:

> *Classroom use of IE strategies can promote both the acquisition and the retention of mechanics-course material well beyond that obtained in traditional practice.*

See Dellwo (2009) for additional commentary on the utility of Hake's gain.

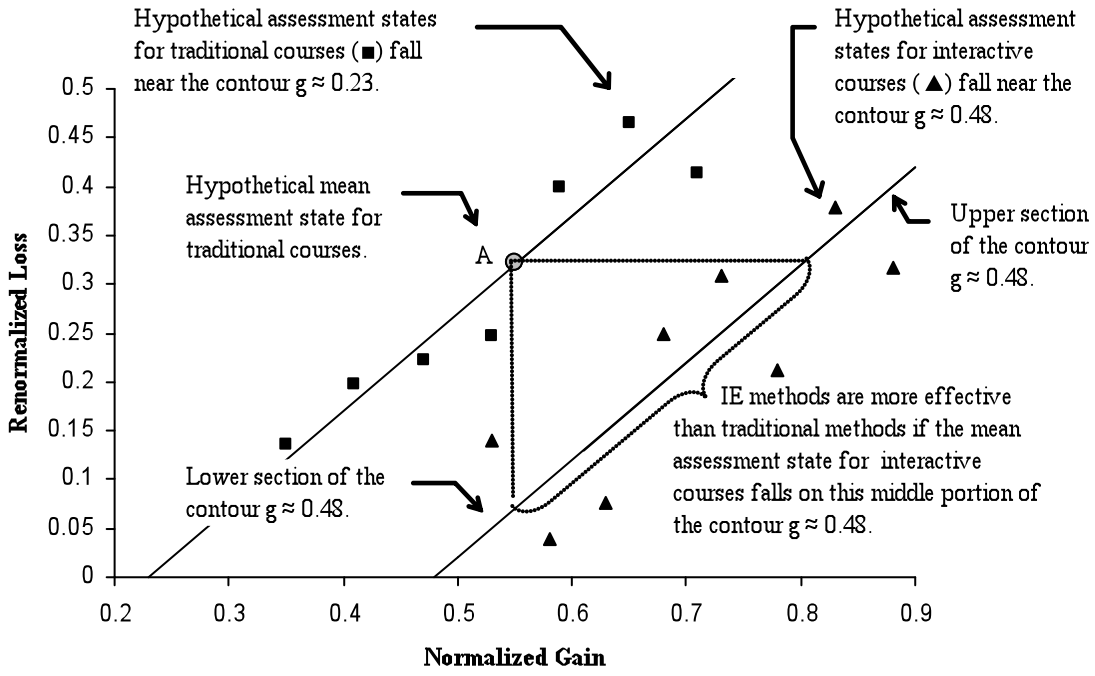

Figure 2. The contours of $g = G - \gamma L$ are parallel lines in the $(G, \gamma L)$ plane. The hypothetical mean traditional state A falls on the contour $g \approx 0.23$ while the mean IE state falls on $g \approx 0.48$.

Disclaimer

The opinions expressed are those of the author and not the U.S. Merchant Marine Academy or the U.S. Department of Transportation.

References

Bond, L. (2005). Carnegie perspectives: Who has the lowest prices? Carnegie Foundation for the Advancement of Teaching, Stanford, CA, retrieved December 21, 2009 from http://www.carnegiefoundation.org/perspectives/sub.asp?key=245&subkey=569.

Cummings, K., Marx, J., Thornton, R., and Kuhl, D. (1999). Evaluating innovations in studio physics. *Physics Education Research, American Journal of Physics,* Suppl. 67 (7), S38-S44.

Dellwo, D. (2000). Assessment of learning in a studio calculus course. *Proceedings CAORF/JSACC 2000 International Multi-Conference on Instructional Technology*, Computer Aided Operations Research Facility (CAORF), United States Merchant Marine Academy, Kings Point, N.Y., July 3-7, D1-1 to D1-15.

Dellwo, D. (2009). Reassessing Hake's Gain. Preprint, available on request.

Ecker, J. (1996a). Studio Calculus. *Ontario College Mathematics Association (OCMA) News and Views* 2, June, 2-3.

Ecker, J. (1996b). *Studio Calculus*, preliminary edition. New York: HarperCollins College Publishers.

Epstein, J. (2007). Development and validation of the calculus concept inventory, in Pugalee, Rogerson, and Schinck, eds. *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community*, September 7-12, retrieved December 21, 2009 from http://math.unipa.it/~grim/21_project/21_charlotte_EpsteinPaperEdit.pdf .

Hake, R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics* 66 (1), 64-74.

Hake, R. (2004). Re: pre-post testing in assessment. Retrieved December 21, 2009 from http://listserv.nd.edu/cgi-bin/wa?A2=ind0408&L=pod&P=R9135&I=-3.

Hake, R. (2006). Possible palliatives for the paralyzing pre/post paranoia that plagues some pep's. *Journal of Multidisciplinary Evaluation* 6, November, available at http://evaluation.wmich.edu/jmde/JMDE_Num006.html.

Hestenes, D., Wells, M., and Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher* 30, 141-158.

Hovland, C., Lumsdaine, A., and Sheffield, F. (1949). The baseline for measurement of percentage change, in C. Hovland, A. Lumsdaine, and F. Sheffield, eds., *Experiments on Mass Communications*. Princeton, N.J.: Princeton University Press, 284-292.

Libarkin, J. and Anderson, S. (2005). Assessment of learning in entry-level geoscience courses: results from the geoscience concept inventory. *Journal of Geoscience Education* 53 (4), 394-401.

McConnell, D., Steer, D., Knott, J., Van Horn, S., Borowski, W., Dick, J., Foos, A., Malone, M., McGrew, H., Greer, L., and Heaney, P. (2006). Using conceptests to assess and improve student conceptual understanding in introductory geoscience courses. *Journal of Geoscience Education* 54 (1), 61-68.

Meltzer, D. (2002). The relationship between mathematics preparation and conceptual learning gains in physics: a possible "hidden variable" in diagnostic pretest scores. *American Journal of Physics* 70 (12), 1259-1268.

Rogosa, D. (1995). Myths and methods: “myths about longitudinal research” plus supplemental questions, in J. M. Gottman, ed., *The Analysis of Change*. Mahwah, N.J.: Lawrence Erlbaum Associates, 3-66.

Ross, S. (2004). *Introduction to Probability and Statistics for Engineers and Scientists*, 3rd ed. New York: Elsevier Academic Press.

Sokoloff, D. and Thornton, R. (1997). Using interactive lecture demonstrations to create an active learning environment. *The Physics Teacher* 35, 340-347.

Suskie, L. (2004a). *Assessing Student Learning: A Common-Sense Guide*. New York: Jossey-Bass, Wiley.

Suskie, L. (2004b). Re: pre-post testing in assessment, retrieved December 28, 2007 from http://lsv.uky.edu/scripts/wa.exe?A2=ind0408&L=assess&T=0&O=A&P7492 or http://tinyurl.com/akz23

Taylor, J. (1982). *An Introduction to Error Analysis*. Mill Valley, CA.: University Science Books.

Willett, J. (1989a). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement* 49, 587-602.

Willett, J. (1989b). Question and answers in the measurement of change, in E.Z. Rothkopf, ed., *Review of Research in Education*. Washington, D.C.: American Education Research Association, vol. 15, 345-422.

Willett, J. (1994). Measurement of change, in T. Husen and T. N. Postlethwaite, eds., *The International Encyopedia of Education*, 2nd ed. Oxford, U.K.: Pergamon Press, 671-678.

Willett, J. (1997). Measurement of change: What individual growth modeling buys you, in E. Amsel and K. A. Renninger, eds., *Change and Development: Issues of Theory, Methods, and Application*. Mahwah, N.J.: Lawrence Erlbaum Associates, 213-243.