Instituut ELAN
Instituut voor
Expertise-ontwikkeling in het VO
Lerarenopleiding
Aansluiting VO-HO
Nascholing in het VO

Universiteit Twente
de ondernemende universiteit

A Pre-test-Corrected Learning Gain

Concept version July 2007

A.B.H. Bos, C. Terlouw (University of Twente, Enschede)
A. Pilot (University of Utrecht, the Netherlands)
(2007)

ELAN doc 2007 – 004
September 2007
GW-ELAN.07.350

Universiteit Twente
GW - ELAN
Postbus 217
7500 AE ENSCHEDE

tel.: 053 - 489 3560
fax: 053 - 489 4755
elan@edte.utwente.nl
http://www.utwente.nl/elan

A Pre-test-Corrected Learning Gain

A.B.H. Bos, C. Terlouw (Twente Technical University, Enschede) & A. Pilot (University of Utrecht, the Netherlands) (2007)

Index

Abstract:

In three different contexts a strong power law relation $y_i = x_i^{1-B}$ was found between the pre-test values x_i and post-test values y_i of individual students, as well as the corresponding relation $\langle y \rangle = \langle x \rangle^{1-B}$ between average group pre-test <x> and average group post-test <y> values. The exponent B in this law is a *pre-test corrected learning gain.* A nominal scale for calculated B-values is suggested. The best method for assessing B is a combination of a plot for visual checking of test data followed by a numerical non linear least squares fit for estimating parameter B and its error. The use of group averages appears to give systematically low B values. It is shown that if pre- and post-test scores are relatively precise, then comparing learning gain exponents has a much higher statistical power than the use of effect sizes representing the post-tests of control and test groups.

1 INTRODUCTION

Post-test only designs may be useful for testing new ideas but for strong tests of hypotheses that rule out alternative interpretations, more sophisticated designs are useful (Shadish, Cook, & Campbell, 2002). Comparing pre and post-test data allow the calculation of *learning gain*, but there is, surprisingly, some reluctance to measure gain, perhaps, as discussed by Hake (Hake, 2006), due in part to the influence of the well known paper “How we should measure ‘change,’ - or should we?” (Cronbach & Furby, 1970).

In a Solomon Four Group Design (S4GD) (Engelenburg van, 1999; Solomon, 1949) two sets of two groups each are formed by a randomization procedure. The first set consists of a control and a test group that is each given a pre-test, the treatment, and the post-test. The second set consists of a control and a test group that is each given the treatment, and the post-test, but no pre-test. The S4GD has two advantages: (1) the pre-test given to the first set gives an indication of the degree of equivalence of the control and test groups after the randomisation, in addition to allowing measurement of the pre-to-post-test gain. If there is a statistical significant difference between the average pre-test scores of the two pre-test groups the whole experiment may be flawed. (2) Because the first set of control and test groups is given a pre-test and the other set is not given a pre-test, the S4GD design makes the *feared* pre-test sensitization (Bos & Terlouw, 2005; Willson & Putnam, 1982) visible.

The outcome of the intervention is in many cases reported as an *effect size* based only upon the difference in post-test scores of the test and control groups. In the psychological literature various effect sizes are defined: e.g., “Δ” defined by (Glass, 1976) [the same as the “effect size” used by Bloom (Bloom, 1984)], and “d” defined by Cohen (Cohen, 1988). Effect sizes generally express the difference between two groups with different treatment in terms of the number of standard deviations. For example Cohen’s d is defined as:

$$d = |m_a - m_b| / [(sd_a^2 + sd_b^2)/2]^{\mathbf{0.5}} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

where m_a and m_b are population means and where the denominator is the root mean square of standard deviations for the A- and B-group means, sometimes called the “pooled standard deviation.”

Cohen suggested that as very rough rule of thumb d = 0.2, 0.5, 0.8 imply respectively “small,” “medium,” and “large” effects. At this point it should be stressed, that *effect size* is not the same as *gain*. For example, in calculating Cohen’s “d”, even if average pre and posttest scores are used so that for $|m_a - m_b|$ in Eq. (1) is the average gain or loss, that average gain or loss is divided by pooled standard deviation. Some indication of *differences in gain* are found in literature, using raw scores of different treatment groups and using a F-test or a t-test on posttest scores. In a report on educational research on a sociology class by Neuman (Neuman, 1989) *performance* is gauged by the score on a multiple choice test and *gain in knowledge* is defined as the difference between the posttest and pretest scores, i.e., for each student gain = (posttest score_i) – (pretest score_i) = $y_i - x_i$. Here the posttest questions are the same as the pretest questions but are rearranged. Neuman found a negative correlation between gain and the pretest score x_i for each student.

Similarly, a negative correlation was found by Hake (Hake, 1998a) between $<y> - <x>$ and $<x>$, where the angle brackets indicate averages over entire classes.

In order to compare the effectiveness of different types of mechanics courses Hake analyzed the results of 62 courses in high schools, colleges, and universities. As a "rough measure" of average effectiveness of a course the average normalized gain $<g>$ for a course was defined as the actual average gain ($<y> – <x>$) divided by the maximum possible average gain ($1 – <x>$), i.e.

$$<g> \; = \; (<y> – <x>) \,/\, (1 - <x>) \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; .(2)$$

where the angle brackets <........> signify averages over entire courses, and scores are normalized so that $0 \leq <y> \leq 1$ and $0 \leq <x> \leq 1$.

On the base of this approach "High-g" courses were categorized as those with $<g> \geq 0.7$ and "Low-g" courses as those with $<g> < 0.3$. The *average normalized* gain was used to demonstrate a nearly two-standard deviation superiority of courses using *“interactive engagement”* methods over those using “*traditional*” methods. Here interactive engagement methods were defined as those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors. As discussed by Hake (Hake, 2002a), it was latter found that the normalized gain had earlier been used independently by Hovland et al. (Hovland, 1949), who called it the “effectiveness index,” and Ghery (Ghery, 1972), who called it the “gap closing parameter”. In recent physics educational research literature the term “normalized gain” is normally employed.

Next to its simple form and its intuitive appeal, $<g>$ is widely used because it compensates to some extent for the variable average pre-test scores in different courses. Thus it can be regarded as a *pre-test corrected learning gain*. An indication that this compensation in fact takes place follows from low correlations between average normalized gain $<g>$ and average pre-test scores $<x>$ (Hake, 1998a, 1998b). In this way it was possible to meaningfully compare the effectiveness of courses with a wide range of average pre-test scores ranging from 18% (a Dutch high school) to about 70% (Harvard).

It should be indicated that Hake also discussed [see e.g., footnote #46 of (Hake, 1998a) and (Hake, 2002a, 2002b)] *single-student* normalized gain, that is, for the ith student,

$$g_i \; = \; (y_i – x_i) \,/\, (1 – x_i) \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; .. \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad . \; .(3)$$

and the two ways of calculating an average normalized gain: (a) from Eq. (2) and (b) from

$$<g> = (\Box_{i = 1 \text{ to } N}\, g_i)/ N \; . \quad (4)$$

where N is the number of students in the class who take both the pretest and the post-test.

The reason that, for $N \geq \sim 20$, Eqs. (2) & (4) usually give $<g>$’s within about 5% of one another are given in footnote #46, and reasons for preferring Eq. (2) to Eq. (4) in educational research are given in (Hake, 2002b).

In this article we suggest on the basis of ecological ("classroom") experiments a model to calculate *pre-test corrected learning gains* from pre- and post-test data in a O_1XO_2 group design, where O_1 is an observation (pre-test), X is some form of treatment and O_2 is an observation after X (the post-test) (Shadish et al., 2002). However, our procedure may be better diagrammed as $OX_1OX_2OX_3$ . . . . where the O’s represent the *same* test given 2 to 6 times in succession (see Table 1 and Fig. 1), and the X’s (treatments) are the computer’s showing the students the correct answer after every missed question.

It must be noted that in the educational literature a pre-test is also considered to be a form of intervention which may have significant influence on the effect of the following intervention, a phenomenon commonly described as pre-test sensitization as discussed in the Introduction (Bos & Terlouw, 2005). In the present work the pre-test was given immediately before the treatment and the post-test was given immediately after the treatment of between 6 and 20 minutes (see Table

1, row 8), so pre-test sensitization is more apt to occur. Our present observations of apparent pre-test sensitization do not necessarily contradict the findings of Henderson (Henderson, 2002) who observed no pre-test sensitization, because in Henderson's analysis (a) the test was designed to measure conceptual understanding (as opposed to declarative knowledge as in the present work) in a conceptually difficult area (introductory physics), and (b) the treatment lasted about a half year, so the time between pre and post-tests was about a half-year rather than a few minutes as in the present work.

In this article the test results of three experiments will be shown to fit a power law relation $y= x^{1-B}$ using matched single student pre and post-test data of a group of students. This same law will be applied to the (Hake, 1998a, 1998b) average *group* data. We shall show that B is a *pre-test corrected learning gain just as is Hake's normalized average gain <g>.* Using Monte Carlo routines, a few methods for estimation of the parameters and parameter errors will be examined. Finally analysis of post-test-only data will be compared to an analysis using the suggested pre-test corrected learning gain in order to compare the statistical power of the two methods.

2 MATERIALS & METHODS

2.1 Tests

In three experiments students following Chemistry, Information Science or French were selected from a population of 194 students of a school for pre university education, average age 17 years. Formative *computer* tests consisted mainly of fill-in-the-blank questions and a few multiple choice questions. The answer to the fill in question was correct if the answer exactly matched the key so the student had to know the answer and to type it correctly. To some questions 2 or 3 keys (= correct answers) were present. The computer was programmed to react to a wrong answer by immediately showing the student the correct answer. The questions were presented in random order, as were the 4 alternative responses to each of the multiple choice questions. The students were allowed to make the test during a 50 minute session as often as the liked. The average result of all tests would contribute in a marginal and only positive way to the end mark of the summative paper and pencil test that was to follow after a few weeks.

Table 1 shows the various parameters that characterize the three experiments A, B, and C of this study.

Experiment	**A**	**B**	**C**
Number of students	32	27	72
% Female	24	26	44
Subject	Information Science	French	Chemistry
# questions	27	40	51
# fill-in-the-blank questions	22	40	49
# multiple choice	5	0	2
Avg. test time (min.)	6.1	7.5	19.7
Tests / student	3.6	4.5	2.0
# Pre / post test pairs	70	94	72

Table 1. Test data for experiments A, B, C.

In experiment A, the students were asked to study at home a 40-page chapter of their textbook on computer science, which some of them did and most of them did not. In experiment B, 40 words were selected from an article in a popular French scientific magazine targeted for youth. In a sentence in French one word was underlined and the contextual translation of the word in correct Dutch was asked. The words were selected using different frequency classes (i.e. common words as well as rare words were asked). The students had no special preparation for this test. For experiment C, a chapter on organic chemistry was assigned as homework prior to the test. The students should have studied the chapter at home, but most of them had not. From the material in the chapter (reaction types, oxidation of simple carbon compounds, industrial synthesis of epoxyethane) were relevant questions were made.

2.2 Simulations, statistical analysis, and numerical integration.

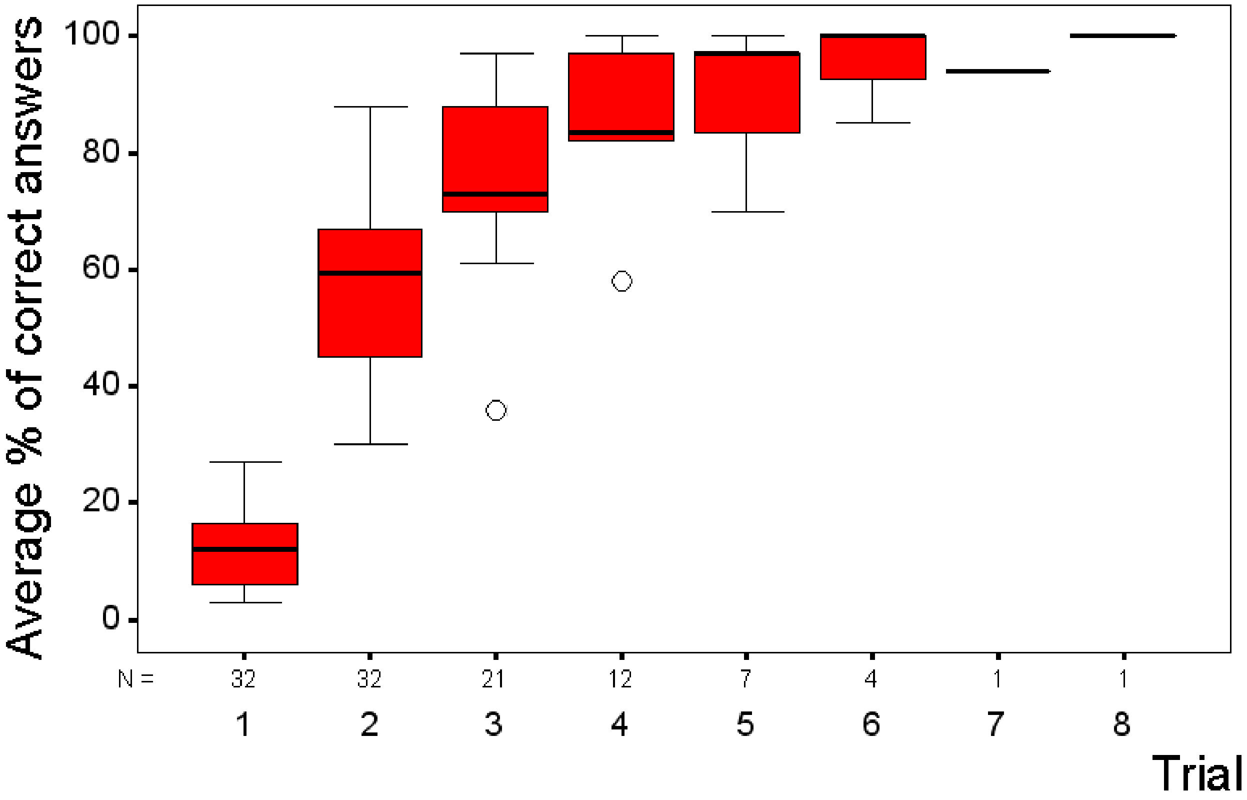
For use in the Monte Carlo procedures a uniform pseudo random generator was coded in C^{++} using Borland's C-Builder version 4. Instead of the built-in 2-byte random library function an algorithm adapted from a subtractive method by Knuth (Knuth, 1981) and making use of 32 bit integer arithmetic was employed. These uniform deviates were transformed to a normally distributed deviate using a algorithm ported from a Pascal routine by (Press, 1989), also based on Knuth. The statistical subroutines written in the C^{++} software were tested by comparing them with results obtained with the statistical package SPSS v.11.01 and the curve fitting program Graphical Analysis 3.1.

Incomplete Beta and Gamma Functions given by the C^{++} software were compared with values given in the Handbook of Mathematical Functions (Abramowitz & Stegun, 1968). Numerical integration was performed with Maple version 7.

3 RESULTS

The average scores for experiment A (in % of max score) are displayed in Fig 1. Each time the test was made the average scores were higher and with each iteration the increase in average scores was smaller.

Fig. 1. Average percentage scores vs trial number for experiment A. N is the number of students engaging in each trial. The Tukey boxplot indicates the minimum and maximum scores, the lower and upper quartiles, the median, and the outliers (circles).



The time needed to complete a test decreases with each iteration. As can be expected from the power law of practice (Anderson, Fincham, & Douglass, 1999), the average time □t (in seconds) needed for completing the test can be described by

$$\square t = \square t_1 * I^{-C} \quad \ldots\ldots\ldots\ldots\ldots \ . \ . \ . \ . \ . \ . \ . \ . \ . \ldots \ . \ . \ (5)$$

where t_1 is the time needed for the test when made for the first time, I the iteration and C is an exponent reflecting the learning rate. Values for □t_1 and C are given in Table 2.

model: t = t_1 . I $^{-C}$	**t_1 (sec.)**	**C**	**R**
A (Information Science)	540	0.56	0.99
B (French)	782	0.68	0.997
C (Chemistry)	1607	0.91	1.0

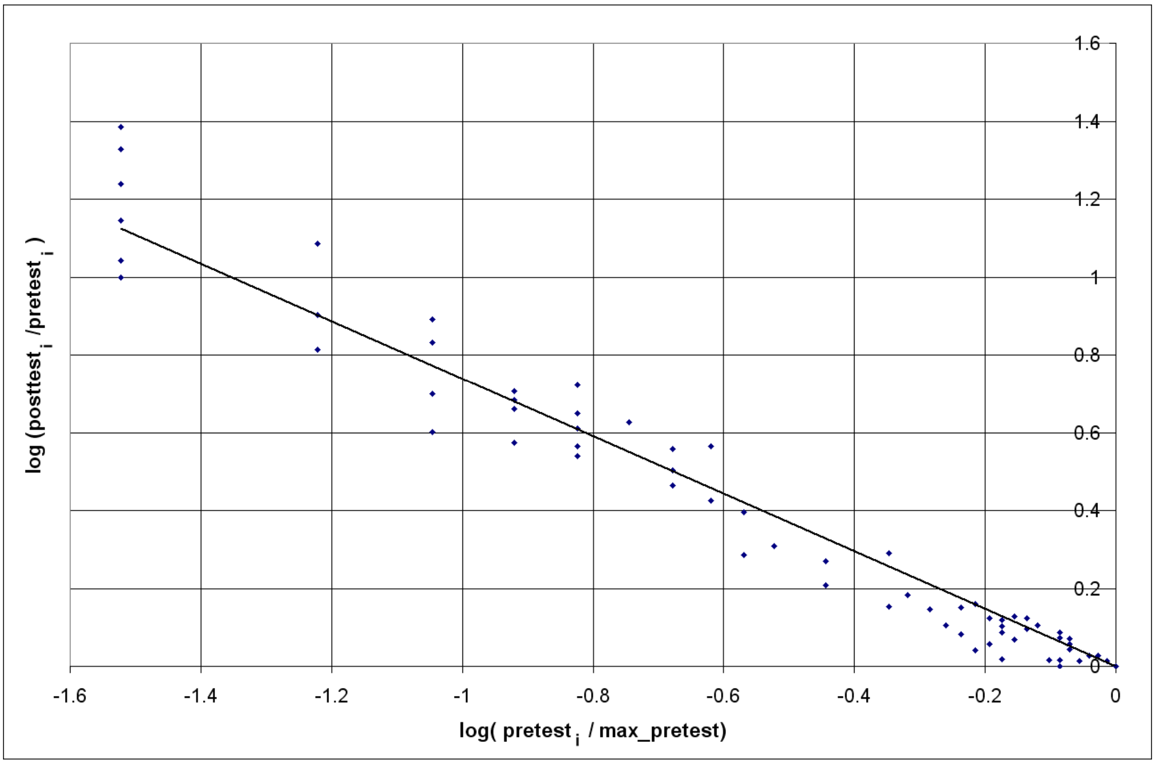
Table 2 : Parameters of the power function giving the time needed for one test at I iterations.

In Table 2 the relatively high exponent “C” in the power function t = t1* I –C for the chemistry test is probably due to the fact that the first time the test was given a problem had to be solved. The second time simple retrieval was sufficient.
When the score of test # n of a particular student is divided by the score of the foregoing test (# n-1) the ratio f = $score_n/score_{n-1}$ is a ***single-student*** growth factor f_i. Test # n-1 can be considered as a pre-test, test # n as a post-test, hence

$$f_i = y_i / x_i \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ldots \ . \ . \ (6)$$

The double logarithmic plot of f_i against x_i of Fig. 2 (for experiment A) shows that f follows a power law f = x $^{-B}$.

Fig. 2. Log f $_i$ = log [(score on test n of student $_i$) / (score on test n of student $_i$ on test n – 1)] = log [post-test$_i$ / pre-test$_i$] on the ordinate is plotted vs log [pre-test$_i$ / maximum_pre-test] = log(x_i) on the abscissa, where f_i is the growth factor (y_i /x_i) for student i.



From Fig. 2 the data indicate: $\log f = -B \log(x) = \log (x^{-B})$ . . …….. (7)

and taking antilogarithms: $f = x^{-B}$ . . . . ….. . . . . …. (8)

and therefore: $B = - \log f / \log x$ …… . .. .(9)

Eq. (6) and Eq. (8) yield: $y = x^{1-B}$ ….. . . . . (10)

Eq. (6) and Eq. (9) yield: $B = - \log (y/x) / (\log x)$. ………….. (11)

From the plot the tangent B applies to the *group* B since all single student data of all students in the group are plotted, but with of Eqs. (9) and (11), a *single-student* B_i can also be calculated individually for each pair of adjacent tests:

$B_i = - \log(f_i) / \log(x_i)$. . . . . (12)

In experiment A the individual B_i's are not dependent on x_i (F-test, p = 0.44). The deviation of the individual B_i's from the average B increases with x_i. The correlation of B_i with x_i is statistically not significant (R = -0.159, p=0.19).

Similar results as in experiment A are found in experiment B and C. A summary is given in Table 3. The correlation coefficient R stems from the regression lines in the double logarithmic plots (such as Fig. 2), and n is the number of data points (such as those in Fig. 2 for experiment A).

Experiment	B	Error in B	R	n
A (Information Science)	0.74	0.015	0.98	70
B (French)	0.68	0.015	0.96	94
C (Chemistry)	0.66	0.012	0.93	72

Table 3. B values for experiments A, B, and C.

3.1 Comparison of methods for estimation of B.

Several methods are available to estimate B. To get some idea about the outcome of different methods of estimating B from pre/post-test data, simulation procedures were invoked.
From a variety of numerical and graphical methods for estimating B the following were chosen:

(a) Calculation the slope from $\log(f_i)$ against $\log(x_i)$ plots (as in Fig. 2) The slope of the regression line has the value -B.

(b) In an iterative numerical procedure the least squares fit of experimental y_i values with x_i^{1-B} a value of B and its error can be found.

(c) A plot of $\log(y_i)$ vs $\log(x_i)$ gives a straight line, in accord with Eq. (10) $y = x^{1-B}$. The slope of the regression line has the value 1-B.

(d) As indicated above in Eq. 9, for each set of data points of a single student (x_i ,y_i), B_i can be calculated using the formula

$B_i = - \log(y_i/x_i)/\log(x_i)$ . . . . (12)

Since the deviation of individual B_i's from the average B increases linearly with x_i, a weighting factor $1/x_i$ is appropriate in averaging B_i's to obtain the group average B.

In a series of Monte Carlo simulations from a hypothetical student population with normally distributed pre-test values and post-test values calculated with the assumed power relation with normally distributed errors ranging from 3 to 30% superimposed on pre and post-test values, B values were calculated. If *robustness* is of a method is defined as giving an accurate and precise estimation of B under varying conditions, it was found that method "a" is the most robust if outlying log f_i and log x_i values (log f_i □ 2 and/or log x_i □ -2) in plots similar to Fig. 2 are neglected. This is because extreme data point values have an extreme influence on the estimated slope in logarithmic plots. Making a plot is highly suitable for visual inspection. If a data point is an outlier and the reason for this is obvious (e.g. a student aborting a test after answering one question) neglection can be considered.

The error in the parameter B is of crucial importance in assessing the statistical significance of differences between experimental outcomes. From the Monte Carlo simulations it was concluded that the most precise (reproducible) and accurate estimation of the **error** in B can be made using method "b". These findings were implemented in a C^{++} -computer program that takes the slope of log(f) versus log(x) as a starting point for an iterative non linear curve fitting, y = x^(1-B) using an Newton-Raphson like first order approximation in order to find the least squares minimum and to calculate the parameter error. If two sets of experimental data, such as that for experiment A, are analyzed by this computer program, the statistical significance p of the difference in the estimated B's is given by a Student-t-test.

3.2 Relation between exponent B and average normalized gain <g> by Hake.

Even if the individual student data are not available for an analysis of B, a *conservative* estimation of B (i.e., estimated B lower than the actual B) is possible using group averages.
An example of the use of group averages is Hake's analysis of data for various traditional and interactive engagement mechanics courses. As indicated in Section I, Eq. (2), Hake defined an average normalized gain <g> as

$$<g> \;=\; (<y> - <x>) \,/\, (1 - <x>) \quad . . . .(2)$$

If it is assumed that Eq. (10), which describes experiments A, B, and C also (at least approximately) applies to the group data analyzed by Hake then Eqs. (2) and (8) yield

$$<g> \;=\; (<y> - <x>)/\,(1 - <x>) = (<x>^{1-B} - <x>) \,/\, (1 - <x>) \; . . . . \;. .(10a)$$

From each pair of randomly chosen <x> , <y> values with <x> < <y> ,
0 < <x> < 1 and <y> □ 1 :
(a) B may be calculated from

$$B = -\log(\, <y> / <x> \,) \,/\, \log(\, <x> \,) \; . . . . (11)$$

and (b) the normalized gain <g> may calculated from equation 2

In Fig. 3 these <g> and B values are displayed by small dots. The relation between B and <g> appears to be somewhat fuzzy, with g □ B.

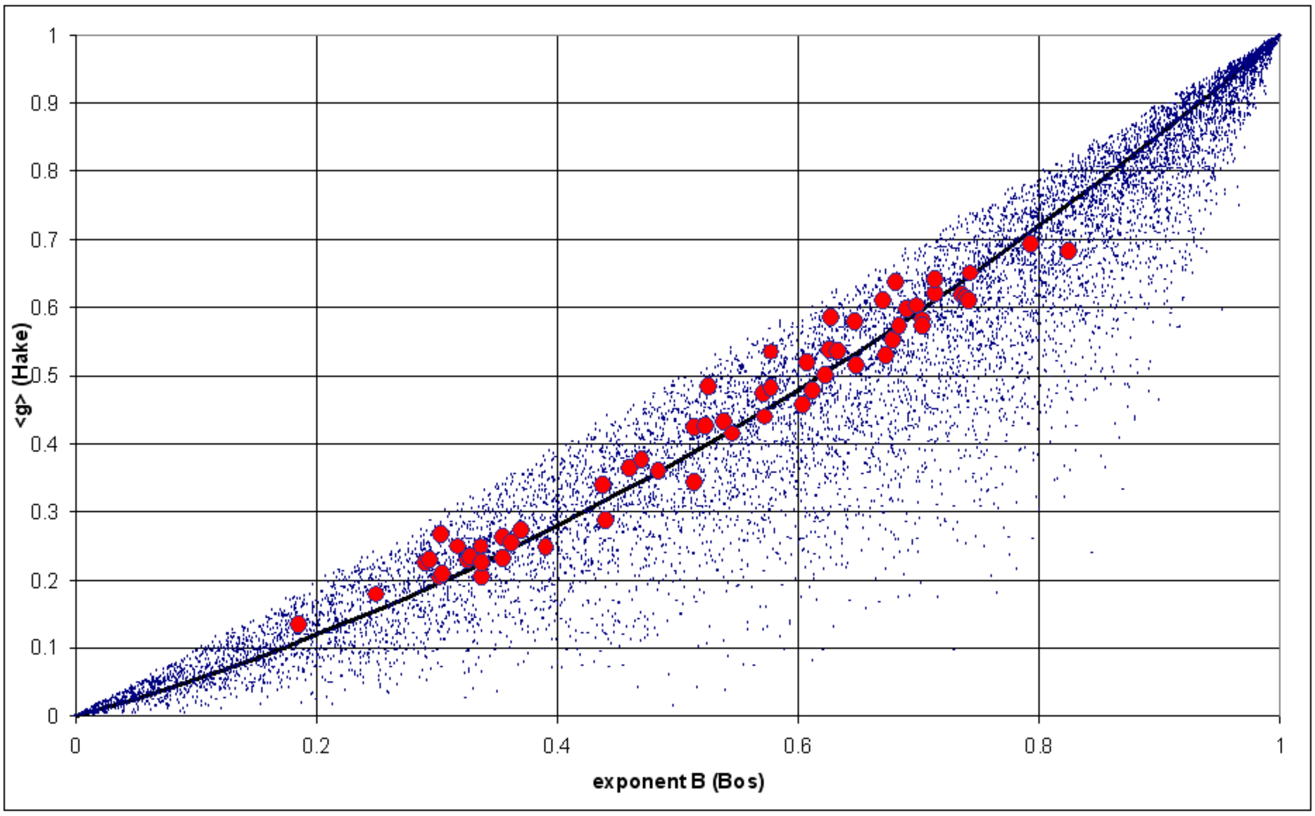


Fig. 3 : <g> (Hake) vs B (Bos) as discussed in the text. The large circles are derived from the <x>, <y> data of Hake (1998a), The curve is found by numerical integration combined with quadratic regression of average <g> values on B (see Table 4).

The average value of <g> (<<g>>) for a certain B can be found by integration of <g> over a definite <x>-interval. In Table 4 average values of <<g>> for B-values over the interval 0 < <x> < 1 are given.

B	**<<g>>**
0	0
0.100	0.0690
0.200	0.1385
0.300	0.2154
0.400	0.2984
0.500	0.3885
0.600	0.4868
0.700	0.5950
0.800	0.7148
0.900	0.8488
1	1

Table 4 : average value of <g> = <<g>> for fixed B-values over the <x> interval between 0 and 1.

By using the data of Table 4 in a quadratic regression of <<g>> on B, it was found that <<g>> can be approximated with the function <<g>> = $0.5B^2$ +0.5B (R = 0.9998). The solid curve in figure 3 is the function <g> = ½B^2 + ½B The (unweighted) coefficient of correlation between the experimental (Hake) data and this function is 0.98.

The estimation of B by using group averages as in Fig. 2 is called *conservative*, because measured B's tend to be systematically smaller than the B's that are used in the Monte Carlo routine with a model again using normally distributed pre-tests values and with normally distributed errors in pre and post-tests. This difference between calculated and actual values of B increases with precision of pre and post-test and is caused by ceiling and floor effects, since the floor effect may increase the apparent pre-test average and the ceiling effect may decrease the apparent post-test average. In both cases lower B-values are calculated.

3.3 A reported correlation between Gain and Pre-test examined.

In identifying factors that influence pre-test scores of an introductory sociology course Neuman used 50 multiple choice questions as pre-test and in a scrambled form as post-test (Neuman, 1989). The *performance* was defined as percentage of correct answers, the *gain* was defined as the difference between pre and post-test percentage of correct answers: *difference*= <%post test> - <%pre-test> . The total group number N was 106 students. Missing data resulted in lower numbers for specific analyses. Average pre and post-test data are given in Table 5.

<%pre-test>	<%post-test>	difference
42.63 ± 8.96 (N=101)	74.15 ± 12.48 (N=104)	31.44 ± 11.60 (N=89)

Table 5: pre/post test data from (Neuman, 1989). *Difference* is (<%post-test> - <%pre-test>)

Neuman reported that the correlation between the gain and pre-test for the 89 students who took both the pre and post-tests was -0.57 from which he concluded that his students learned more when they enter the course knowing less.

From the group data in Table 5 a learning gain exponent B can be estimated from Eq. (11):

B = - [log (<y>/<x>)] / (log <x>) = -log(0.7415/0.4263)/log(0.4263) = 0.65.

For a careful calculation of B only the matched data from the 89 students who took both the pre and post-test ought to be used. Careful pre/post testers use only matched data sets as discussed in Hake (Hake, 1998b, 2002a, 2002b).

With this approximate B-value and 100 normally distributed pre-test values with same mean and standard deviation as in table 5, post-test values were calculated in a Monte Carlo routine using equation 10 : $y = x^{1-B}$. The same mean and standard deviation for calculated post-test data and *difference* (between pre and post-test) as in the Neuman report were found in this simulation. The correlation coefficient between pre-test and *difference* was practically the same as in the Neuman study (R = -0.57) if an error of 10% upon pre and post-test values is superimposed, as shown in Table 6.

error in pre/post test (%)	R
0	-0.60 ± 0.063 (N=100)
10	-0.58 ± 0.063 (N=100)
30	-0.54 ± 0.065 (N=100)

Table 6: Correlation R between pre-test and difference (= <%post-test>-<%pre-test>)

3.4 Power considerations using classical analysis versus Learning Gain Exponent calculations.

In order to get some idea of the power of the classical method of comparing post-test scores only in terms of *effect sizes* versus determination of B in the power law $y= x^{1-B}$ (method 'b') as proposed in this article several experimental situations were simulated. One of these experiments will be described here in detail: two sets of post-test scores with group sizes N = 28 were gener-

ated using normally distributed pre-test scores with mean 0.50 and standard deviation 0.15. Post-test scores corresponding to B = 0.66 were used for one set, and post-test scores corresponding to B = 0.55 were used for the other set. Upon the pre and post-test and B-values normally distributed errors were superimposed and the statistical significance of the differences of the outcomes between the two sets were calculated. The relative errors in pre and post-tests were equal and varied between 0 and 30%. A error of 1% on the B values was set. The post-test scores of both data sets were compared using a double sided Student t-test ("classical method"). Also the gain exponents B and the parameter errors were calculated with our software and compared with a double sided Student t-test. If p was higher than 0.05 a type II error was indicated.
This procedure was repeated 10000 times giving an indication of type II error frequency, denoted by in the literature. The power of a test is defined as 1-. The power and the statistical significance level are strongly interdependent. Usual values are = 0.05 and 1- is 0.80 (Dupont & Plummer, 1998). In Fig. 4 the power is given as a function of the error in pre/post-test (in %). As can be expected the power decreases with increasing error.

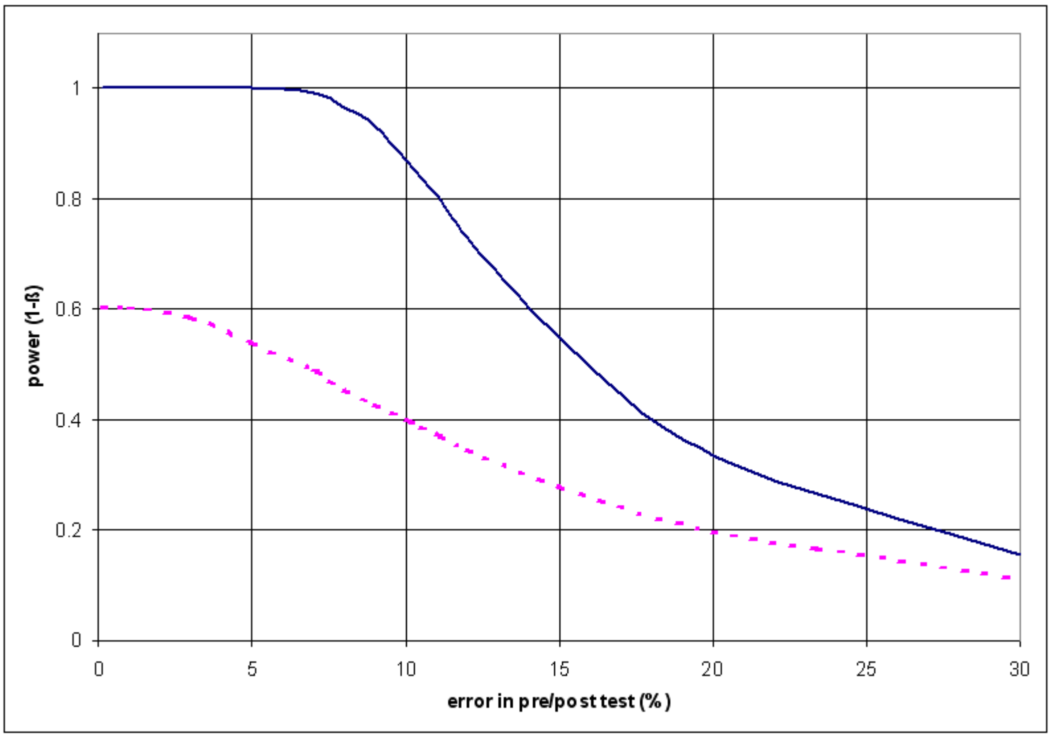


Figure 4. Power as a function of error in pre and post-test. Upper curve : estimation and comparison of B-values. Dotted curve : comparing post-test data only (*classical* method).

In the classical method using differences between post-test scores only in this particular case the power of 0.80 is never attained. In the method of calculating and comparing B's the power is higher than 0.80 if the errors in pre and post-test do not exceed 11%. In all other similar trials similar results were found : even with small number of students (e.g. N=10) the power of the method described in this article using pre and post-test data approached unity where the classical method in which only post-test scores are compared the power was far below the acceptable value of 0.80.

4 CONCLUSIONS AND DISCUSSION

In three very different contexts (Chemistry, French, and Information Science) in comparable experimental settings we found a strong, definite power law relation $y = x^{1-B}$ between pre-test x and post-test y values of individual students. The parameter B can also be derived from (a) using average post-tests <y> and pre-tests <x> or (b) by averaging B_i's of individual students as determined from $y_i = x_i^{1-B}$. The fact B's derived by using single student data and group averages agree with one another suggests analysis can be carried out even when single student data are not available. A serious flaw is the systematic decrease of accuracy caused by bottom and ceiling effects especially using group averages. On the basis of our experience with the type of testing described in this paper and taking into consideration Hake's findings, a suggestion for a nominal scale of pre-test corrected learning gains could be as follows (Table 7):

exponent B	**gain**
B ≤0.40	low
0.40 < B < 0.60	average
B ≥ 0.60	high

Table 7 : classification of pre-test corrected learning gains.

Though a strong power law relation between pre and post-test results in very different contexts was found, it must be emphasized that in other experimental settings other relations may prove to be appropriate, although in cognitive psychology power laws mostly with time (latency) as independent variable, but with other variables as well, seem to be ubiquitous (Ritter & Schooler, 2002). Another relation in other settings may prove to give a better fit, but the purpose of the model stays the same: elucidating deeper relations or impact of different types of interventions.
Log(y/x) against log x gives the opportunity for a rapid inspection of the data. The noise in pre and post-test is damped. If some aberration of normal test procedure happens (e.g. a student aborting the test) these data are spotted easily and could be inspected. For error calculations we prefer the fitting of data (x,y) to the function $f(x) = x^{(1-B)}$, without any transformations except normalizing the x and y values in the interval (0,1).
A combination of the visual inspection of the log(y/x) against log x followed by the least squares curve fitting procedure for estimation of B and the error in B is advised. The OXO-setting (pre-test/treatment/post-test) in combination with gain calculations (including estimating parameter errors) seems to have a very high statistical power even with small numbers of students. Crucial is the availability of precise and accurate tests otherwise the GIGO-effect is encountered, a classical effect in applied information science (garbage in, garbage out). Very small differences in learning gains can be traced giving the opportunity to evaluate subtle effects.

The authors are greatly indebted to Richard R. Hake; Indiana University Emeritus for thorough and fundamental review of this article.

References

Abramowitz, M., & Stegun, I. A. (1968). *Handbook of Mathematical Functions* (5th ed.). New York: Dover Publ. Inc.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and Retention: A Unifying Analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1120-1136.
<http://act-r.psy.cmu.edu/papers/133/PraRet1999.pdf> (3.1 MB).

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Bos, A. B. H., & Terlouw, C. (2005). *Met ICT gevoelig maken voor het leren van bètabegrippen.* Paper presented at the Meten en Onderwijskundig Onderzoek, Proceedings van de 32e Onderwijs Research Dagen 2005, Gent.
[The effect of pre-test sensitizing in a digital system on the acquisition of science concepts.]

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"- or should we ? *Psychological Bulletin, 74*(1), 68-80.

Dupont, W. D., & Plummer, W. D. (1998). Power and Sample Size Calculations for Studies Involving Linear Regression. *Controlled Clinical Trials 1998, 19*, 589-601.
"PS", an interactive program for performing power and sample size calculations is available at <http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize>.

Engelenburg van, G. (1999). *Statistical Analysis for the Solomon Four-Group Design*. Enschede: Twente Univ. Faculty of Educational Science and Technology.
<http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/10/91/76.pdf>. (548 kB).

Ghery, F. W. (1972). Does Mathematics Matter? In A. Welch (Ed.), *Research Papers in Economic Education* (pp. 142-157): Joint Council on Economic Education.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Hake, R. R. (1998a). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys., 66*, 64-74.
online at <http://www.physics.indiana.edu/~sdi/ajpv3i.pdf> (84 kB).

Hake, R. R. (1998b). *Interactive-engagement methods in introductory mechanics courses*, from http://www.physics.indiana.edu/~sdi/IEM-2b.pdf

Hake, R. R. (2002a). *Assessment of Physics Teaching Methods.* Paper presented at the Proceedings of the UNESCO-ASPEN Workshop on Active Learning in Physics, Univ. of Peradeniya, Sri Lanka.
online at <http://www.physics.indiana.edu/~hake/Hake-SriLanka-Assessb.pdf> (84 kB).

Hake, R. R. (2002b). Lessons from the physics education reform effort. *Ecology and Society, 5*(2), 28.
online at <http://www.ecologyandsociety.org/vol5/iss2/art28/>

Hake, R. R. (2006). Possible Palliatives for the Paralyzing Pre/Post Paranoia that Plagues Some PEP's" [PEP's = Psychometricians, Education specialists, and Psychologists]. *Journal of MultiDisciplinary Evaluation, 6*, 59-71.
online at <http://evaluation.wmich.edu/jmde/JMDE_Num006.html>.

Henderson, C. K. (2002). Common Concerns about the Force Concept Inventory. *The Physics Teacher, 40*(9), 542-547.
<http://homepages.wmich.edu/~chenders/Publications/TPT2002.pdf> (408 KB).

Hovland, C. I. (1949). A Baseline for Measurement of Percentage Change. In P. F. Lazarsfeld & M. Rosenberg (Eds.), *The Language of Social Research: a Reader in the Methodology of Social Research.* (pp. 77-82): Free Press.

Knuth, D. E. (1981). Seminumerical Algorithms. In *The Art of Computer Programming* (2nd ed., Vol. 2, pp. 116 ff). Reading, Mass.: Addison-Wesley.

Neuman, W. L. (1989). Which Students Learn the Most and Why? A Replication and Extension of the Szafran Pretest Study. *Teaching Sociology, 17*(1), 19-27.
An ERIC abstract is online at <http://tinyurl.com/2szfx5>.

Press, W. H. (1989). Numerical Recipes in Pascal. In *The Art of Scientific Computing* (1st ed., pp. 213-226). Cambridge: Univ. Press.

Ritter, F. E., & Schooler, L. J. (2002). The learning curve. In *International encyclopedia of the social and behavioral sciences.* (pp. 8602-8605). Amsterdam: Pergamon.
online at <http://ritter.ist.psu.edu/papers/ritterS01.pdf> (156 kB)

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*, 137-150.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*(2), 249-258.
An ERIC abstract is online at <http://tinyurl.com/3aqa9u>