The discriminative learning gain: a two-parameter quantification of the difference in learning success between courses

Julie Direnga, Dion Timmermann, Ferdinand Kieckhäfer, Andrea Brose & Christian Kautz

To cite this article: Julie Direnga, Dion Timmermann, Ferdinand Kieckhäfer, Andrea Brose & Christian Kautz (2018): The discriminative learning gain: a two-parameter quantification of the difference in learning success between courses, Australasian Journal of Engineering Education, DOI: 10.1080/22054952.2018.1553357

To link to this article: https://doi.org/10.1080/22054952.2018.1553357

Published online: 11 Dec 2018.

Submit your article to this journal

View Crossmark data
The discriminative learning gain: a two-parameter quantification of the difference in learning success between courses

Julie Direnga\textsuperscript{a}, Dion Timmermann\textsuperscript{b}, Ferdinand Kieckhäfer\textsuperscript{c}, Andrea Brose\textsuperscript{b} and Christian Kautz\textsuperscript{a}

\textsuperscript{a}Engineering Education Research Group, Hamburg University of Technology, Hamburg, Germany; \textsuperscript{b}Center for Teaching and Learning, Hamburg University of Technology, Hamburg, Germany

1. Introduction

How can we measure what students have learnt in a specific course? Researchers in STEM Education Research often use pre- and post-tests to assess the effect of a course or intervention on student learning (e.g. Hake 1998). The pre-test thereby provides a baseline measurement to account for a possible difference in levels of prior knowledge between the groups. The post-test gives insight on the absolute level of achievement at the end of the period of interest.

Sometimes, non-identical pre- and post-tests (NIPPs) must be used, i.e. the pre-test and the post-test are different instruments. This might be the case when an instrument valid for use as both pre- and post-test does not exist (see for example Direnga et al. 2014). Consequently, the result of the analysis is not a learning gain as used by Hake and others.

The use and evaluation of NIPPs has not yet become a standard in the discipline-based education communities. While the established evaluation methods for identical pre- and post-tests (IPPs) have been thoroughly discussed (e.g. Bond 2005; Cronbach and Furby 1970; Hake 2006; Marx and Cummings 2007), evaluation methods for NIPPs have not been addressed to the same extent by the community. Analysis of covariance (e.g. Engqvist 2005) and multiple linear regression (e.g. Theobald, Freeman, and Stone 2014) have been proposed, but are often not applicable as will be explained below.

This article summarises and extends earlier results published in Direnga et al. (2014) and Direnga et al. (2017). We first discuss shortcomings of IPPs and argue why using NIPPs is often a better option. After laying out our theoretical framework, we present an overview of established methods for evaluating pre- and post-test data. We then introduce the discriminative learning gain (DLG) based on linear regression with confidence bounds as a tool to quantify as well as easily visualise the difference in courses with respect to their performance on NIPPs. Additionally, we show how to apply an effect size measure to the regression model. Reducing the complexity of the data is necessary to effectively quantify differences in learning success between courses. While the test scores contain information at the individual student level, applying a regression reduces the complexity of the data to two parameters which describe course performance. The interpretation of those parameters will be discussed in detail below. A comparison of the methods is made.

2. Motivation for using non-identical pre- and post-tests

There are two inherent shortcomings with validity when using IPPs:
Many introductory engineering courses introduce new concepts and vocabulary. In order to successfully test student understanding of the concepts taught, the test instruments must often make use of this newly introduced vocabulary. If the test-takers cannot understand what the questions are asking, the interpretation of test scores cannot be as valid as intended. Unfortunately, this makes it impossible to give a pre-test that measures student understanding of the concepts in the same way a post-test would measure it.

(2) The score range of any test instrument is finite. In using test instruments, one must keep in mind that strong occurrences of floor or ceiling effects have a negative impact on the validity of test score interpretations. Score distributions that are heavy on either one of the range extremes should therefore be avoided by choosing appropriate instruments. Due to instruction between the tests, an improvement in understanding is expected. On IPPs, this would be represented by a shift in score distribution towards the higher end. Especially for large effects of instruction, this can lead to the problem that the real shift in understanding cannot be displayed by the limited score range.

For these reasons, we argue that using NIPPs is often a better option.

3. Theoretical framework

The statistical method we call DLG can be applied to many types of data. The purpose for which we want to use and promote it is to evaluate effectiveness or quality of teaching. Using some sort of gain or change measure is quite common in Physics and Engineering Education Research. Von Korff et al. (2016) conclude that gain is a ‘powerful tool’ for the purpose of comparing classes. General criticism of measuring score change has been expressed for example by Cronbach and Furby (1970). They argue that although gain or change measures are widely used they are not appropriate in most cases. Their argument is based on the idea that the true score is unknown due to measurement errors. The article ‘How We Should Measure “Change” – Or Should We?’ was commented by Bond (2005), Hake (2006), and many others. In Hake’s opinion, this ‘pre/post paranoia’ is one of the reasons that results from education research often do not have an impact on reforms. In this article, measurement errors are assumed to be random, so that scores are assumed to be non-biased.

Conceptual understanding can be seen as a proxy for effectiveness or quality of teaching because it is a critical achievement on the way from novice to expert thinking (Mazur 1997). Furthermore, we see conceptual understanding as a construct that can be measured by means of validated instruments such as concept inventories (CIs). Therefore, we can measure and quantify the quality of teaching by comparing multiple choice test results and CI scores (Adams and Wieman 2011). In doing so, we do not focus on individuals but on aggregated data from entire cohorts.

4. Established methods

In this section, we will give an overview of some established methods for evaluating learning success: the average normalised gain ($g$), normalised change ($c$), and analysis of covariance (ANCOVA). An overview of similarities and differences among these methods and the DLG can be found in Table 2 in the Discussion section.

4.1. Average normalised gain

A very widely known study that uses learning gains is the study by Hake (1998) comparing interactive-engagement and traditional teaching in various physics classes. The average normalised gain, denoted by $g$, is a well-established measure for assessing data on teaching effectiveness. It is defined as follows (Hake 1998):

$$g = \frac{\text{absolute gain}}{\text{maximum possible gain}}$$

$$= \frac{(\text{post})_{\text{ave}} - (\text{pre})_{\text{ave}}}{100\% - (\text{pre})_{\text{ave}}}. \quad (1)$$

The absolute gain is the gain of class average pre- and post-test scores, and $(\text{pre})_{\text{ave}}$ and $(\text{post})_{\text{ave}}$ are the class average scores of the pre- and post-tests, respectively, given as a percentage of the maximum possible score.

Lines of constant $g$ can be visualised by plotting the absolute gain vs. the average pre-test score, as shown in Figure 1. The range of $g$ starts at the horizontal ($g = 0$) and reaches its maximum of $g = 1$ at a slope of $-1$. The data displayed in Figure 1 were taken from Hake (1998) for illustration purposes. Each data point displayed corresponds to one course. If linearity of the measure is assumed, it can be seen that the average $g$ of the interactive engagement courses is slightly more than twice as high as that of the traditional courses. Hake’s study and the quantification of learning through $g$ give instructors the opportunity to compare their achieved gains to these values.

It should be noted that the special case of $\text{pre} = 100\%$ is not mathematically defined by...
Equation (1), but since this case is highly unlikely due to several factors (operation on averages, good test construction) and easy to interpret (no more gain possible), there is no need for further elaboration. Still, there have been critical voices by Marx and Cummings (2007) concerning the definition of $g$. Based on the reasonable assumption that the performance on an identical test will improve after instruction, $g$ was designed for positive gains, only. Although it would not be mathematically incorrect to apply Equation (1) for post < pre, $g$-values in the range $\frac{|\text{post}\text{-pre}|}{\text{pre}} < 1$ can be attained, which makes averaging more difficult. Also, the interpretation is questionable when relating a loss to the maximum possible gain. A loss at an already low initial score would be considered less negative than the same absolute loss at a higher initial score, which is contrary to the intent of the average normalised gain. Marx and Cummings (2007) proposed a different definition for negative values, which is presented in the following section.

4.2. Normalised change

To make $g$ applicable to negative gains in an easy-to-interpret manner, the normalised change described by Marx and Cummings (2007) can be applied. While a positive absolute gain is related to the maximum possible gain as in Equation (1), a negative absolute gain is related to the maximum possible loss, i.e.:

$$c = \begin{cases} \frac{\text{post}-\text{pre}}{100\%} & \text{if } \text{post} \geq \text{pre} \\ \frac{\text{post}-\text{pre}}{\text{pre}} & \text{if } \text{post} < \text{pre} \end{cases}$$

Note that Marx and Cummings (2007) advised to first calculate the gains of individual students and then average the gains. This procedure requires the data from pre- and post-tests to be linked. However, they also state that 'for large numbers of students the numerical difference is small' between averaging the scores and averaging the gains. Under these circumstances, $c$ could still be applied to unlinked data without having to expect larger errors.

The diagram shown in Figure 2 can be seen as an extension of Figure 1. The data were collected using NIPPs (for details see Direnga et al. 2014). All $c$ are negative, indicating that the post-test was more difficult than the pre-test. It can be seen that the normalised change in the interactive engagement courses, $c_{IE} = -0.06$, is greater than in the traditional courses with $c_{T} = -0.30$. This corresponds to the results found by Hake (1998).

Even though this definition allows a reasonable interpretation of negative gains and comparisons between gains, these are only valid when comparing gains from the same pre-/post-test combination. This is due to the fact that a gain of zero cannot be interpreted as 'no learning', as the pre- and post-test had a different level of difficulty. Therefore, the data shown in Figure 1 is explicitly not displayed here together with the NIPP data as this would suggest that a comparison between these datasets was valid.

However, it is feasible to compare the differences in average gains $\Delta c$ for any two courses, i.e. for interactive engagement and traditional courses:

$$\Delta c_{IE\text{-}T} = c_{IE} - c_{T}. \quad (3)$$

It is worth noting that if Equation (3) is applied to the data displayed in Figure 1 (Hake's data from IPPs) and Figure 2 (our data from NIPPs), respectively, these two
values do not substantially differ from each other. For the data from Figure 1, we get $\Delta c_{IE-T} = 0.48 - 0.23 = 0.25$, while for the data from Figure 2, we get $\Delta c_{IE-T} = -0.06 - (-0.30) = 0.24$. Assuming that instruction using interactive engagement does result in a greater learning effect and that the employed NIPP combination does measure this effect, this similarity supports the hypothesis that the normalised change is a valid statistical method for assessing NIPP data.

One disadvantage shared by $g$ and $c$ is a bias in favour of high pre-test populations (Nissen et al. 2018).

### 4.3. Analysis of covariance (ANCOVA)

In non-randomised studies, differences in treatment groups with respect to concomitant variables (here pre-test scores) can obscure the visibility of the treatment effect on the outcome variable under investigation (here post-test scores). ANCOVA has proven to be a powerful and widely used tool to account for such variables. The model for a single-factor ANCOVA with fixed treatment effect can be written as

$$Y_{iq} = \bar{Y} + \tau_q + \gamma (X_{iq} - \bar{X}) + \epsilon_{iq},$$

where

- $X_{iq}$ and $Y_{iq}$ are the observed pre- and post-test scores of the $i$-th individual on the $q$-th treatment,
- $\bar{X}$ and $\bar{Y}$ are the mean pre- and post-test scores over all individuals and treatment groups,
- $\tau_q$ are fixed treatment effects with $\sum \tau_q = 0$,
- $\gamma$ is a regression coefficient denoting the effect of $X$ on $Y$ (i.e. the slope), and
- $\epsilon_{iq}$ are assumed to be independent and normally distributed errors (Neter, Wasserman, and Kutner 1985).

It becomes evident that the regression coefficient representing the slope, $\gamma$, is assumed to be independent of the treatment $q$. This assumption is not always adequate, and a failure to recognise when it fails can lead to a misinterpretation of the results (see e.g. Engqvist (2005) for details). We must therefore conclude that ‘when the treatment regression lines interact with the concomitant variable in the form of nonparallel slopes, covariance analysis is not appropriate.’ (Neter, Wasserman, and Kutner 1985, 851).

### 5. Discriminative learning gain

After having presented an overview of some established methods, we propose the DLG as a tool to compare the effectiveness of teaching based on pre- and post-test score pairs. When ANCOVA fails due to non-parallel slopes, Neter, Wasserman, and Kutner (1985) suggest that ‘separate treatment regression lines should be estimated and then compared’. The DLG is based on a regression line with confidence bounds. Here, the individual pre-test scores $X_i$ are the independent variable and the individual post-test scores $Y_i$ are the dependent variable. As the regression lines are estimated separately, the index $q$ indicating the different treatment groups will be neglected for better readability. An example of the raw data, namely the distribution of pre- and post-test scores for one treatment group, is shown in Figure 3.

Even though the variation is rather large in the individual data, we found that the mean post-test scores $\bar{Y}$ at each possible pre-test score level $X_i$ are quite well described by a linear model (see Figure 4). As mentioned above, we only want to make inferences about the entire cohort instead of individuals.

![Figure 3. Distribution of individual score pairs on non-identical pre- and post-tests, N = 828.](image)

![Figure 4. Mean post-test scores for each pre-test score and resulting regression line. (Multiple occurrences of identical score pairs are not represented here, see Figure 3 instead.)](image)
Therefore, the linear model is appropriate even for large variation in individual data.

5.1. Regression line

The linear model is given by

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad (5) \]

The line parameters of the model are estimated using an ordinary least squares approach by minimising the following term with respect to the variables \( \beta_0 \) and \( \beta_1 \):

\[ Q = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2. \quad (6) \]

The resulting line parameters \( b_0 \) and \( b_1 \) are called point estimators of \( \beta_0 \) and \( \beta_1 \):

\[ b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})}, \quad (7a) \]

\[ b_0 = \bar{Y} - b_1 \bar{X}, \quad (7b) \]

where \( \bar{X} \) and \( \bar{Y} \) are the overall average pre- and post-test scores, respectively. This results in the estimated points on the line

\[ \hat{Y}_h = b_0 + b_1 X_h. \quad (8) \]

Here, the index \( h \) denotes that the variables \( X_h \) and \( \hat{Y}_h \) are continuous, whereas \( i \) and \( j \) stand for the individual observations at discrete test scores and distinct pre-test score levels, respectively.

5.2. How to interpret the DLG regression line

The line is best interpreted as the expected post-test score for ‘model students’ with the respective pre-test scores, given that linearity is assumed and the best fit to the data is applied. The DLG provides two valuable pieces of information:

1. The general post-instruction level is represented by \( \hat{Y}_{50%} \) the estimated post-test score at a pre-test score of 50 %, which is equivalent to the average value of the line.
2. The discriminative effect of the instruction for students with different pre-test score levels is represented by the slope \( b_1 \). It shows how subgroups of students with different pre-instruction levels responded to the instruction.

A high general post-instruction level and a non-negative slope are of course favourable. A large positive slope would show that the stronger model student benefits a lot more from instruction than the weaker model student, while a negative slope would indicate that either the tests might not be valid in a way that they do not measure the intended criteria, or the instruction created confusion which rather affected students with good conceptual thinking. A slope close to zero represents an equalising effect of instruction. Whether the ideal course should have a strong positive slope (i.e. students with a significantly higher pre-test score should also have a significantly higher post-test score compared to low scoring students) or not (i.e. students should reach more or less the same level after instruction independent of their pre-test scores) is rather a matter of the personal view on education.

What the DLG generally cannot do (just like the established methods) is predict individual students’ post-test scores from their pre-test scores because of the often large variations in individual scores (see Figure 3). Furthermore, NIPPs generally do not allow to identify the pre- and post-test score pairs that are the equivalent to ‘no learning’. Also, the resulting line parameters depend strongly on the particular pre- and post-tests used. To illustrate this dependency, suppose a group was randomly divided into two subgroups. They take the same post-test but different pre-tests. Because the two subgroups were formed randomly, they have the same ability such that one subgroup scores high on the easier pre-test while the other subgroup scores low on the more difficult one. For the same reason, the score distribution on the post-test should not differ between the two subgroups. This can be illustrated as a horizontal shift of the data in Figure 3 and consequently in Figure 4. The general post-instruction level of the subgroup taking the easier pre-test will therefore be lower in comparison to the other subgroup, which would be an incorrect conclusion. Therefore, a single regression line does not tell us much about the effectiveness of the instruction. Instead, it must be compared to other data sets using the same test pair. When interpreting the results, statistical uncertainties have to be considered, e.g. by determining the confidence bounds around the regression lines, as shown in the next section.

5.3. Confidence bounds

As we are dealing with sampled data, the resulting regression parameters, both slope \( b_1 \) and intercept \( b_0 \), can only be estimated with a certain confidence \( 1 - \alpha \). By using a joint estimation technique, we make use of the fact that \( \beta_0 \) and \( \beta_1 \) are not independent. The inequality

\[ N(b_0 - \beta_0)^2 + 2 \sum X_i (b_1 - \beta_1) + \sum X_i^2 (b_1 - \beta_1)^2 \leq F(1 - \alpha; 2, N - 2) \quad (9) \]

describes an elliptical \( 1 - \alpha \) confidence region in the \( \beta_0, \beta_1 \)-space (Neter, Wasserman, and Kutner 1985, 147–148). Here, MSE is the mean squared error.
\[
\text{MSE} = \frac{\sum (Y_i - \hat{Y}_i)^2}{N - 2},
\] (10)

- \(N\) is the sample size,
- \(\alpha\) is the confidence coefficient, and
- \(F(1 - \alpha; 2, N - 2)\) denotes the inverse of the \(F\)-distribution with 2 degrees of freedom in the numerator and \(N - 2\) degrees of freedom in the denominator for the percentile \(1 - \alpha\).

All \(\beta_0, \beta_1\)-pairs contained in this region determine the set of possible regression lines. From this, a hyperbolically shaped confidence band around the estimated regression line delimited by the upper and lower confidence bounds can be derived:

\[
\hat{y}_{\text{upper,lower}} = \hat{Y}_h \pm W \cdot |s(\hat{Y}_h)|.\] (11)

\(\hat{Y}_h\) is defined by Equation (8), \(W\) is a constant given by

\[
W^2 = 2F(1 - \alpha; 2, N - 2),
\] (12)

and the standard deviation of the estimated points is

\[
s(\hat{Y}_h) = \sqrt{\text{MSE} \cdot \frac{1}{N} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}\] (13)

(Neter, Wasserman, and Kutner 1985, 154 and 74). Note that the values of the variance for any \(h\) also depend on data from other pre-test scores.

The regression line parameters \(b_0\) and \(b_1\) given by (7) can also be obtained by performing a weighted least squares approach on the means (see e.g. Neter, Wasserman, and Kutner 1985, 167). In this case, the weights are given by the frequency of each pre-test score \(X_h\). In Direnga et al. (2014) and Direnga et al. (2017), we have focussed on this equivalent approach referred to as weighted linear regression (WLR). The confidence bounds, on the other hand, must be calculated based on the individual data. By taking into account only the weighted means, we would lose valuable information about the confidence because the number of points that the confidence bounds would be based on would never exceed the total number of distinct pre-test scores. This would generally result in wider bounds than necessary.

5.4. How to interpret the DLG confidence bounds

If the experiment were repeated again and again with an assumed constant population, then 95% of the times (for \(\alpha = 0.05\), the entire regression line would lie within the confidence band. We can thus conclude that if the regression line of one data set lies outside of the confidence bounds of the other data set, it is highly unlikely that the data stem from the same population. In other words, we are most likely observing an effect of the intervention to the experiment group.

5.5. Effect size applied to the DLG

The ongoing practice of reporting statistical significance has been heavily criticised for decades (Coe 2002; Cohen 1994; Johnson 1999). Effect size measures (and confidence bounds) are often proposed instead. Correctly interpreted, the \(p\)-value only informs us about the probability of collecting this or more extreme data from our assumed population. It does not report the size of an effect. Furthermore, statistical significance depends heavily on the sample size. Effect size, on the other hand, is a statistic that quantifies the effect of, for example, an intervention and is independent of sample size. It is generally defined as the relation between difference in group means and the common standard deviation (e.g. Cohen 1988, 20). To be conservative, we take the standard deviation of the group (A or B) with the greater variance. Applying this concept to the DLG results in the following pre-test score-dependent expression:

\[
d_{A,B}(X) = \frac{\hat{Y}_A(X) - \hat{Y}_B(X)}{\max(\tilde{\sigma}_A, \tilde{\sigma}_B)},\] (14)

where the calculation of \(\tilde{\sigma}\) is based on the sum of squared errors, i.e. the deviation of individual measurements from the regression line:

\[
\tilde{\sigma} = \sqrt{\frac{\text{SSE}}{N - 2}} = \sqrt{\frac{\sum i \sum j(Y_{ij} - \hat{Y}_{ij})^2}{N - 2}}.\] (15)

The result is again a line and it may be reported as such. In case single values are preferred, we suggest to report \(d_{A,B}(50\%)\). The discriminative effect can be considered by also reporting the effect sizes for the low- and high-achieving groups, e.g. \(d_{A,B}(30\%)\) and \(d_{A,B}(70\%)\). Note that this form of calculating the effect size accounts for the correlation between the pre- and the post-test data and the result is therefore larger than an independent form of \(d\) (Nissen et al. 2018).

5.6. Quantifying the degree of linearity

As shown above, assuming a linear model results in two parameters for each course, which are still easily interpretable. A more complex model might result in a better fit, but also makes the interpretation less intuitive. Therefore, assuming a linear model seems to be the best option.

If we want to test this assumption when applied to specific data, a graphical analysis is often sufficient. There might be occasions, however, when a test for linearity is called for. In this case, the \(F\)-test for lack of fit can be applied.3 One requirement for the application of the lack-of-fit sum of squares (SSLF) is that there are replicates, i.e. multiple observations of post-
test scores for at least one pre-test score. Replicates are very likely to occur in pre-/post-test settings, given that the ratio of number of participants to number of distinct pre-test scores is reasonably high.

The sum of squares due to error can be partitioned into the sum of squares due to pure error plus the sum of squares due to lack of fit:

$$SSE = SSPE + SSLF \quad (16)$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{k} n_j (\bar{Y}_j - \hat{Y}_j)^2, \quad (17)$$

where

- \(k\) is the total number of distinct pre-test scores,
- \(n_j\) is the number of observations corresponding to the \(j\)-th pre-test score,
- \(Y_{ij}\) is the \(i\)-th post-test score observation corresponding to the \(j\)-th pre-test score,
- \(\bar{Y}_j\) are the means over all \(n_j\) post-test score observations corresponding to the \(j\)-th pre-test score (see Figure 4), and
- \(\hat{Y}_j\) is the estimated value at the \(j\)-th pre-test score.

The pure error is thus taking into account the deviations of the individual observations from the means, i.e. the ‘spread’ of the individual data (see Figure 3). The lack of fit, on the other hand, is considering the deviations of the means from the estimated regression line (see Figure 4), weighted by the number of observations of each pre-test score, which is a good estimate for the degree of linearity of the data.

We can then test our null hypothesis \(H_0\), i.e. that a linear model is adequate, by comparing the test statistic \(F^*\) to the value of an \(F\)-distribution.\(^4\)

$$F^* = \frac{MSLF}{MSPE} = \frac{\frac{SSLF}{k-2}}{\frac{SSPE}{N-k}} \quad (18)$$

$$F^* \leq F(1-\alpha; k-2, N-k) \rightarrow H_0, \quad (19)$$

where MSLF and MSPE stands for lack of fit mean square and pure error mean square, respectively (Neter, Wasserman, and Kutner 1985, 127–130). If the null hypothesis cannot be rejected, we can conclude that a non-linear model is not more adequate than the linear one.

6. Demonstration

To demonstrate the DLG, we first apply the \(F\)-test for lack of fit to the demonstration data shown in Figures 3 and 4. We then calculate the regression lines with confidence bounds for this and two other data sets. One set was artificially generated for illustrative purposes. The real data stem from several cohorts of a first-year university course on statics for engineers. The administered tests were the Force Concept Inventory (Hestenes, Wells, and Swackhamer 1992) as pre-test and the Concept Assessment Tool for Statics (Steif and Dantzer 2005) as post-test.

6.1 Lack-of-fit test

The Force Concept Inventory consists of 30 items which theoretically would result in \(k = 31\) possible pre-test scores. However, the extreme scores \(X_{1j} = 0\) and \(X_{31} = 30\) did not occur in the data (see Figures 3 and 4). With \(k = 29, N = 828,\) and \(\alpha = 0.05\), we obtain

$$F^* = \frac{13.56}{18.74} = 0.72 < 1.50. \quad (20)$$

Hence, linearity can be assumed for this data. This result supports the impression from the graphical analysis.

6.2 Treatment 1 vs. treatment 2

Figure 5 shows the resulting regression lines for two cohorts labelled treatment 1 and treatment 2. The maximum possible scores on the tests are reflected by the ranges of the axes: 30 points on the pre-test and 27 points on the post-test. The shaded areas represent the respective 95 % confidence bands calculated by Equation (11). Compared to the variation in individual test scores (see Figures 3 and 4) the band of treatment 1 seems quite narrow. This is because it does not serve as a prediction for the next individual measurement, but for the next regression line in case the experiment was repeated and a new sample of the same size was generated from the same population. Treatment 1 results in a higher general post-instruction level \(\hat{Y}_{50\%1} = 12.5\) compared to treatment 2, where \(\hat{Y}_{50\%2} = 8.4\). Furthermore, treatment 1 shows a greater discriminative effect.

Figure 5. Comparing treatment 1 and treatment 2 with 95 % confidence bounds on the regression lines.
with a slope of $b_{1,1} = 0.48$ compared to the slope of treatment 2, $b_{1,2} = 0.33$. The confidence bands in Figure 5 do not overlap except for the very low pre-test scores between 0 and 2 points where very few measurements exist.

Both estimated regression lines do not penetrate the other confidence band at any value on the pre-test score range. At a pre-test score of 50%, we have a gap of 4.1 points between the means of the two treatments, 3.1 points between the inner confidence bounds, and 4.9 points between the outer confidence bounds. Because we have to account for two uncertainties of $\alpha$, the total confidence is $(1 - \alpha)^2$. For $1 - \alpha = 0.95$, this results in a total confidence of 90.3% that both regression lines lie within the confidence bands. If a larger total confidence is required, $1 - \alpha$ must be chosen accordingly. The resulting interpretation would be that with 90.3% confidence, a model student in treatment 1 with a pre-test score of 15 points will score between 3.1 and 4.9 points higher on the post-test compared to the same model student in treatment 2.

The ANCOVA shows that the interaction term between treatment group and pre-test score is statistically significant at $p < 0.001$, i.e. the slopes are different. We can conclude that the discrimination effect of treatment 1 is significantly higher than that of treatment 2. In other words, while treatment 1 results in superior expected post-test scores for all model students independent of pre-test score, students with higher pre-test scores tend to benefit even more from treatment 1 compared to students with lower pre-test scores. Finally, calculating the effect size based on Equations (14) and (15) yields quite large values:

$$d_{1,2}([30\%, 50\%, 70\%]) = [0.73, 0.94, 1.14].$$

The established methods also show a difference between the treatments in terms of general post-instruction level, but the difference in discriminative effect is obscured (see Table 1). In case of $g$ and $c$, the negative values for each data set can be misleading ($g_1 = -0.04$, $g_2 = -0.38$; $c_1 = -0.06$, $c_2 = -0.35$).

### 6.3 Treatment 1 vs. treatment 3

Another example is presented in Figure 6 which shows a comparison of treatment 1 to a third (artificial) treatment. The difference in the sample sizes is represented in the different widths of their confidence bands. For the pre-test score range between $X_{15} = 12$ and $X_{20} = 19$, where the line of treatment 1 intersects the confidence band of treatment 3, we cannot detect a difference between the treatments. The general post-instruction levels hardly differ with $\hat{Y}_{50\%1} = 12.5$ and $\hat{Y}_{50\%3} = 12.6$, but the slopes indicate a strong difference in discriminative effect with $b_{1,1} = 0.48$ and $b_{1,3} = 0.05$. Treatment 3 seems to bring all students to the same level regardless of their pre-test score. The effect sizes are

$$d_{1,3}([30\%, 50\%, 70\%]) = [-0.55, -0.03, 0.48],$$

indicating that stronger students benefit more from treatment 1 and weaker students from treatment 3. The established methods disagree: $g$ shows a very small, $c$ and ANCOVA show no significant difference between the treatments (see Table 1). Again, the strong difference in discriminative effect is obscured by incorrectly assuming equal slopes in the ANCOVA or by ignoring the influence of pre-test score altogether when applying $g$ or $c$.

---

**Table 1**. Results from the comparative analysis of treatments using the different methods. All methods lead to the conclusion that treatment 1 is superior to treatment 2. Similarly, as long as only the first parameter of the DLG (the general post-instruction level) is considered, all methods lead to the conclusion that there is no difference between treatments 1 and 3. Only by inspection of the DLG’s second parameter, the strong difference in discriminative effect becomes visible, suggesting different treatments for students at high and low pretest levels.

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1 vs. 2</th>
<th>Treatment 1 vs. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
<td>$g_1 - g_2 = 0.34$</td>
<td>$g_1 - g_3 = 0.0$</td>
</tr>
<tr>
<td>$c$</td>
<td>$c_1 - c_2 = 0.29$</td>
<td>$c_1 - c_3 = 0.01$</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>$Y_{50%1} - Y_{50%2} = 3.97$ common slope $= 0.43$</td>
<td>$Y_{50%1} - Y_{50%3} = 0.00$ common slope $= 0.42$</td>
</tr>
<tr>
<td>DLG</td>
<td>$Y_{50%1} - Y_{50%2} = 4.1b_{1,1} - b_{1,2} = 0.15d_{1,2} = [0.73, 0.94, 1.14]$</td>
<td>$Y_{50%1} - Y_{50%3} = 0.1b_{1,1} - b_{1,3} = 0.43d_{1,3} = [-0.55, -0.03, 0.48]$</td>
</tr>
</tbody>
</table>
7. Limitations

Statistical tests and models generally have requirements for the data that they can be used for. These requirements will be discussed in this section.

7.1 Under which circumstances can the DLG be used?

The DLG requires the test pair to be the same for each course to be comparable. It is possible, however, to compare the effect sizes \( d_{A,B} \) to \( d_{C,D} \) even if the common test pair of courses A and B is different from the one used in courses C and D.

Another prerequisite is that the DLG works on linked pre- and post-tests of individual students (see Direnga et al. 2016 for a proposal of how to link test scores). The linking procedure must be taken care of when administering the tests.

In general, working with large data sets is favourable as larger sample sizes provide higher confidence. Our experience is that upwards of 50 students, the DLG can be used. However, to achieve reasonably narrow confidence bounds, 150 or more students are desirable.

The model in Equation (5) assumes that all \( Y_i \) are independent normal random variables, with mean \( E(Y_i) = \beta_0 + \beta_1 X_i \) and constant variance. Furthermore, the error terms \( \epsilon_i \) are independent and normally distributed around zero (Neter, Wasserman, and Kutner 1985, 49). Since the score ranges are finite, possible ceiling or floor effects always pose a threat to the normality assumption. If it is violated, the DLG should be interpreted with caution.

7.2. Which tests can be used with the DLG?

The DLG is most robust if the pre-test data has a reasonable spread. Consequently, the pre-test used should have a good discrimination. Pre-tests with a very high or very low difficulty, i.e. ones where most students are at the upper or lower end of the scale causing a floor or ceiling effect are not ideal. Similarly, the post-test should not have a floor or ceiling effect.

When evaluating teaching in engineering mechanics, we obtained data that was readily interpretable with the DLG using the Force Concept Inventory (Hestenes, Wells, and Swackhamer 1992) as pre-test and the Concept Assessment Tool for Statics\(^5\) (Steif and Dantzler 2005) as post-test.

7.3. Can the DLG be applied to identical pre- and post-tests?

The method can also be applied to the special case of IPPs. However, if the effect to be measured is large, there will be a large difference between the pre- and the post-test scores. This makes it more likely for one of the two tests to have a floor or ceiling effect. Figure 7 shows the distribution of pre- and post-test FCI scores on an IPP and Figure 8 shows the respective regression line. In contrast to NIPPs, the line indicating ‘no learning’ can be drawn. Consequently, with IPPs it is possible to make inferences about the course performance from a single regression line by comparing it to ‘no learning’.

Also in contrast to the data from NIPPs in Figure 3, we can observe a ceiling effect on the post-test which is problematic due to the reasons laid out above. One consequence is a reduced estimated slope, especially when there is a non-negligible amount of observations in the higher pre-test score range. The estimated slope of the regression line will often be less than 1 because students in the high pre-test score range do not have many scores to gain. The consequence is an under-estimation of both, the discriminative effect as well as the general post-instruction level.

8. Discussion and conclusion

Based on the argument that using pre- and post-tests that are non-identical is often a better option than using identical tests, we presented several established methods for evaluating pre- and post-test data. After discussing their shortcomings, we introduced the DLG, a simple linear regression with confidence bounds. We demonstrated how one can interpret the line parameters and how to make inferences about differences in learning success between courses using the same NIPPs.

Especially when the differences in learning gains are small, confidence bounds on the regression lines help by giving insight how likely the lines, and therefore the effectiveness of treatments or courses differ. The discriminative effect as an additional output of
the DLG can furthermore reveal differences in treatment effects on different subgroups that might be obscured by scalar learning gains. Calculating the effect size at multiple pre-test score values allows us to quantify this effect.

Table 2 provides an overview illustrating the similarities and differences among the discussed methods. Note that although $g$ and $c$ are not intended for NIPPs, this use case is also considered for illustrative purposes. As mentioned above, $g$ is no longer interpretable in a meaningful way if the post-test score is lower than the pre-test score. The apparent advantage of $g$ and $c$, that courses can be compared regardless of the chosen test, is actually more a matter of IPPs or NIPPs than it is of the method, because the comparability requires the existence of a true neutral element: while this ‘no learning’ reference is generally unknown in case of NIPPs, the zero in $g$ or $c$ can be more easily misinterpreted as such than the line ‘post = pre’ in case of the regression methods ANCOVA and DLG.

While $g$ and $c$ are scalar, the DLG-method returns two parameters for each line: the general post-instruction level and the discriminative effect. Comparing two data sets, the DLG hence returns four, while ANCOVA requires a common slope and therefore returns only three parameters. One disadvantage of a two-parameter learning gain is that it requires a larger sample size to produce reliable results. Whereas $g$ operates on class averages, $c$ allows to consider statistical uncertainty e.g. by calculating the standard error of the mean, although it is not always reported. The regression methods consider statistical uncertainties e.g. by $F$-test, confidence bounds, and effect size. For $g$, the output itself is linear in the sense that a course with $g = 0.4$ achieved twice as much of what was still possible to learn compared to a course with $g = 0.2$. For $c$, linearity is only given on either side of zero. For ANCOVA and DLG, only the output difference to e.g. a reference course is linear (much like a temperature scale).

All presented methods consider only a single covariate, the pre-test score. If more concomitant variables need to be included in the analysis, multiple linear regression (e.g. Theobald, Freeman, and Stone 2014) might be more appropriate. Note that the multiple linear regression model assumes equal slopes like ANCOVA and that with multiple dimensions, the aspect of visualisation is lost.

Considering all these aspects, we conclude that the DLG is a very valuable tool that provides us with two parameters to characterise the effect of a course or treatment in comparison to others. The data can furthermore be easily visualised and the confidence bounds add extra value to the interpretation of the data. While we would generally suggest to look at a research question through multiple lenses, the form of analysis must match the data. This article presents one form of quantitative analysis and compares it to others, enabling users to make informed decisions about which form of analysis is most appropriate in their case. Engineering instructors and researchers who wish to assess their teaching by means of a pair of pre- and post-tests that are not necessarily the same instrument or who want to use another predictor instead of a pre-test can use the DLG as a statistical tool to effectively evaluate, interpret, and compare their data.

### Notes

1. Although exams are often created by highly experienced instructors, they generally cannot provide the same standard of validity.
2. This is especially true for class averages. The chances of obtaining negative gains are higher if the measure is applied to individual scores.

**Table 2.** Illustration of similarities and differences among the methods. The plus and minus signs indicate whether the respective criterium in the first column is (+) or is not (−) fulfilled, and whether there is a high risk of misinterpretation (−−).

<table>
<thead>
<tr>
<th></th>
<th>$g$</th>
<th>$c$</th>
<th>ANCOVA</th>
<th>DLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows for post &lt; pre</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Comparability across instruments</td>
<td>IPP: +</td>
<td>IPP: +</td>
<td>IPP: +</td>
<td>IPP: +</td>
</tr>
<tr>
<td>NIPP: −−</td>
<td>NIPP: −−</td>
<td>NIPP: −</td>
<td>NIPP: −</td>
<td></td>
</tr>
<tr>
<td># of output parameters</td>
<td>One data set: 1 Two data sets: 2</td>
<td>One data set: 1 Two data sets: 2</td>
<td>One data set: 2 Two data sets: 3</td>
<td>One data set: 2 Two data sets: 4</td>
</tr>
<tr>
<td>Considers statistical uncertainty</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Linear</td>
<td>+</td>
<td>+/−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

**Figure 8.** Mean post-test scores for each pre-test score and resulting regression line.
3. if the following assumptions are met: for each pre-test score, the corresponding observations of the post-test scores are independent, normally distributed and have the same variance (Neter, Wasserman, and Kutner 1985, 123).
4. Note that the α in Equation (19) has the same meaning as in Equation (9) but their values can be chosen independently.
5. formerly known as Statics Concept Inventory.
6. Coe (2002, 4) presents multiple ways to interpret these numbers.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) [grant number 01PL11047]. Any opinions expressed here are those of the authors.

Notes on contributors

Julie Direnga is a doctoral student and research assistant in the Engineering Education Research Group at Hamburg University of Technology (TUHH) and an engineer with degrees in General Engineering Science (B.Sc.) and Mechatronics (M.Sc.). Her current research focuses on the assessment of conceptual understanding in introductory mechanics courses.

Dion Timmermann is an electrical engineer (M.Sc.) and educational researcher. He is currently pursuing a Ph.D. in university-level engineering education research at Hamburg University of Technology. His research focuses on students’ conceptual understanding of electrical engineering and the design of instructional materials fostering conceptual change.

Ferdinand Kieckhäfer is a doctoral student and research assistant in the Engineering Education Research Group at Hamburg University of Technology with degrees in General Engineering Science (B.Sc.) and Electrical Engineering (M.Sc.). His areas of research are conceptual understanding in control engineering and the development of concept inventories.

Andrea Brose is a mathematician (Ph.D.) and the head of the Center for Teaching and Learning at Hamburg University of Technology (TUHH), where she coordinates the professional development of faculty and teaching staff in engineering. Previously, she worked as a post-doc in the Engineering Education Research Group at TUHH, where she helped develop and evaluate of evidence-based teaching and learning materials.

Christian Kautz is the head of the Engineering Education Research Group at Hamburg University of Technology. A physicist by training, he holds a degree in university-level physics education research (Ph.D.) from the University of Washington. After three further years in the US, as an assistant professor in the Department of Physics at Syracuse University, he returned to Germany. Besides teaching seminars on engineering education and teaching various introductory engineering courses, he carries out research on student understanding of core concepts in these subjects, as well as on more general cognitive skills.

References


