A statistical method for assessing teaching effectiveness based on non-identical pre- and post-tests

Direnga, J.[1]
Research Assistant, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany

Timmermann, D.
Research Assistant, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany

Brose, A.
Center for Teaching and Learning
Hamburg University of Technology
Hamburg, Germany

Kautz, C.
Professor for Engineering Education, Engineering Education Research Group
Hamburg University of Technology
Hamburg, Germany

Conference Topic: Educational Research Methods

INTRODUCTION

Diagnostic tests and concept inventories are widely used in science and engineering education to monitor the effectiveness of teaching. Often, identical tests are given before and after instruction, allowing direct comparison of respective scores [1]. However, many introductory engineering courses introduce new concepts and vocabulary. In order to test student understanding of the concepts taught, it is often necessary to use vocabulary that is introduced in the course. 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.

In this paper, we will present different measures to assess teaching effectiveness in the cases of (a) identical pre- and post-tests (IPP), and (b) non-identical pre- and post-tests (NIPP). By the latter, we mean that the questions in the pre- and post-tests are completely different.

To illustrate these measures, we will show data for an introductory statics course, which has been taught over the course of eight years using different teaching methods. Two of the discussed measures can be found in the literature [2, 3], while the third is proposed by our group. Using our set of data, we illustrate the advantages and disadvantages of the measures. Finally, we will compare the new method with those found in the literature.

1 DESCRIPTION OF THE DATA

1.1 Course description

To illustrate the different measures presented in this paper, we will use an eight-year dataset from an introductory course on statics held at Hamburg University of Technology. In the academic year 2005 we

[1] Corresponding Author

started to monitor student understanding of concepts in this course. In 2009, interactive engagement (IE)-elements were included in the course.

Apart from the introduction of the IE elements, some organisational changes were made to the course over the time span considered here. In the first four years, statics was covered in half a semester with 180 minutes of lecture and 90 minutes of recitation per week. Since 2009, statics is taught in one semester with 90 minutes of lecture and 45 minutes of traditional recitation per week. In the academic years 2007 and 2008, the course was offered to students in their second semester, while it was offered as a first-semester course in all other years. Two different lecturers held the course, using the same syllabus and lecture notes. An overview of these aspects can be found in *Table 1*.

1.2 Interactive engagement course interventions

The IE methods introduced were tutorial worksheets and Just-in-Time Teaching (JiTT) [4]. The tutorial worksheets are structured group worksheets that are based on research on student understanding. They were developed by our group and modelled after the *Tutorials in Introductory Physics* [5]. Their development and evaluation has been described previously in [6]. To evaluate the effectiveness of their introduction, we used pre- and post-tests.

1.3 Reasons for using non-identical pre- and post-tests

As post-test at the end of the semester, the Statics Concept Inventory (SCI)[2], developed in 2005 by Steif and Dantzler [7] was the obvious choice since data indicate that it “offers reliable and valid measures of conceptual knowledge in statics” [7]. Unfortunately, the SCI is not suitable as a pre-test since it uses many technical terms unknown to students at the beginning of the course. If the SCI were used as a pre-test here, its results would be hard to interpret, as most students would not even understand the questions. Instead, we chose the Force Concept Inventory (FCI) as pre-test at the beginning of the semester. This concept inventory was developed by Hestenes et al. in 1992 to “probe and assess the commonsense beliefs of [...] students” with respect to Newtonian mechanics and to reveal common misconceptions [8]. Hake described the test as ”understandable to the novice who has never taken a physics course, while at the same time rigorous enough for the initiate” [2]. The vocabulary used in the test mostly consists of every-day words and phrases and the problem set-ups involve objects such as cars accelerating, elevators moving, or boxes being pushed. Thus, the problem set-ups in it can be understood by students, even if they have not yet attended a lecture on statics or physics. Although there have been some critical voices to what the FCI actually measures [9], the authors have achieved to refute most of the arguments [10].

While the FCI is a good choice for a pre-test, its use as a post-test for this course would be far from ideal. Since the course focuses on statics, an improved understanding of Newton’s Laws would only be a secondary effect and not a good measure for the success of the course. In summary, the post-test should not be used as a pre-test and the pre-test should not be used as a post-test, but both tests are good choices as pre- and post-test, respectively.

1.4 Example data

Table 1 contains the average pre- and post-test results of the described courses employing FCI as pre-test and SCI as post-test. Judging by the post-tests results, the students who attended the course with tutorials (and possibly JiTT), had a better conceptual understanding of statics than those that only had traditional teaching. One might suspect that this is due to a higher entry-level knowledge of those students. The pre-test results, however, suggest that this is not the case. In our opinion, this clearly shows that the introduction of Tutorials (and also JiTT) had a positive effect on the students’ conceptual understanding of statics gained in the course. But comparing these pre-test and post-test data is not as intuitive, as e. g. the analysis presented in [2]. Therefore, our goal in this paper is to present easy to interpret measures to use in the case of NIPPs.

[2] The current version of the SCI is known as CATS - Concept Assessment Tool for Statics.

			Average Test Score (%)			
Academic Year[a]	Semester	Instructor	Pre-Test	Post-Test	N	Teaching
2005	first	A	53	44	136	(T) Traditional
2006	first	B	51	36	61	(T) Traditional
2007	second	A	51	34	123	(T) Traditional
2008	second	B	50	30	227	(T) Traditional
Weighted Mean			51	35		
2009	first	B	49	44	305	(IE) Traditional + Tutorials
2010	first	B	44	42	369	(IE) Traditional + Tutorials
2011	first	A	49	46	465	(IE) Traditional + Tutorials + JiTT
2012	first	A	42	42	363	(IE) Traditional + Tutorials + JiTT
Weighted Mean			46	44		

Table 1: Overview of changes in the course over a span of eight years, including pre- and post-test scores.

[a] An academic year always starts on first of October and ends on the last of September, and thus spans two calendar years. Given here is the calendar year the academic year started in.

2 ASSESSMENT METHODS FOR TEACHING EFFECTIVENESS

In order to state whether the teaching effectiveness of one course is greater than that of another, one should not simply judge by exam results. There are two main uncertainties. Was the students' pre-instruction level of knowledge the same for both courses? And if different exams are given in course A and course B, how do we make sure they are comparably difficult? Of course, instructors do their best to create fair, valid, and reliable exams each year. Nevertheless, they are never tested as rigorously as concept inventories such as the FCI [8] or SCI [7], which have been around for many years. Also, exams could hardly be administered as pre-tests, which are necessary to measure the pre-instruction level. But even if data from pre- and post-tests exist, there remains the question of how to relate post-instruction level to pre-instruction level, in order to assess teaching effectiveness.

What do we expect from a good measure? We think that the following criteria are desirable. The measure should

i) likewise be *applicable to NIPPs* and IPPs,
ii) *allow comparison* of data from different tests (in the case of IPPs) or from different pairs of pre- and post-tests (in the case of NIPPs),
iii) have a *neutral element*, e.g. zero, representing "no learning",
iv) be interpretable as a *linear* measure of teaching effectiveness, and
v) be bounded, i.e. the range of the values is finite.

For our purpose, i) is not only desirable but necessary. In this section, established methods as well as a new method proposed by us are discussed in detail, pointing out their respective advantages and disadvantages.

2.1 Average normalised gain

In [2], Hake concluded that IE teaching methods show a larger effectiveness than traditional (T) methods. To support this claim, he introduced the so-called "average normalised gain (ANG)"[3], which is defined "as the ratio of the actual average gain $\langle G\rangle$ to the maximum possible average gain" [2]:

$$\langle g\rangle = \frac{\%\langle G\rangle}{\%\langle G\rangle_{max}} = \frac{\%\langle S_f\rangle - \%\langle S_i\rangle}{100 - \%\langle S_i\rangle}. \tag{1}$$

[3] spelling has been adapted to British English

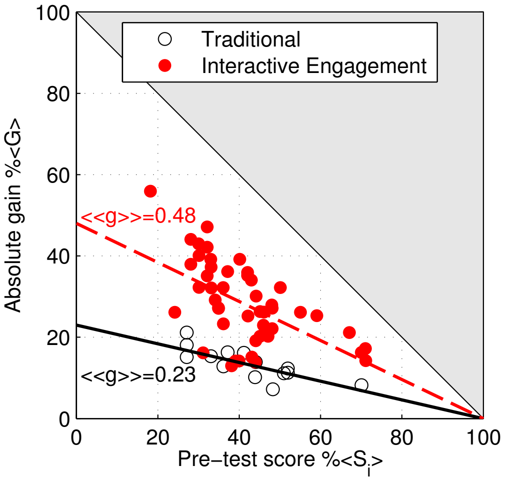


Fig. 1: Illustration of the average normalised gain based on data from [2].

Using the same notation as in [2], $\langle G\rangle$ is the gain of class average pre- and post-test scores, and $\%\langle S_i\rangle$ and $\%\langle S_f\rangle$ are the class average scores of the pre- and post-tests, respectively, given as a percentage of the maximum possible score. Note that the special case of $\%\langle S_i\rangle = 100$ is not mathematically defined by *Eq.* (1). However, 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.

This ANG is a well established measure for assessing data on teaching effectiveness. Lines of constant $\langle g\rangle$ can be visualised by plotting the absolute gain $\langle\%G\rangle$ vs. the average pre-test score $\%\langle S_i\rangle$, as shown in *Fig. 1*. The range of $\langle g\rangle$ starts at the horizontal ($\langle g\rangle = 0$) and the value increases with increasing inclination angle, reaching its maximum of $\langle g\rangle = 1$ at a slope of -1. The data displayed in *Fig. 1* were taken from [2] for illustration purposes. If linearity of the measure is assumed, it can be seen that the average of the average gains $\langle\langle g\rangle\rangle$ of the IE-courses is slightly more than twice as high as that of the T-courses.

There have been critical voices concerning the definition of $\langle g\rangle$. In [3], Dellwo argues that *Eq.* (1) cannot "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" and proposes to decompose the measure into a normalised gain and a renormalised loss. In order to monitor these quantities, the decomposed normalised gain operates at the item level of the pre- and post-tests. When applying NIPPs, it is impossible to match test items to see whether knowledge has been acquired, retained or lost which is why Dellwo's critique is irrelevant in this case.

Based on the reasonable assumption that the performance on an identical test will improve after instruction[4], the definition stated above was designed for positive gains, only. Although it would not be mathematically incorrect to apply *Eq.* (1) for $\langle S_f\rangle < \langle S_i\rangle$, $\langle g\rangle$-values in the range $[-\infty, 0)$ can be attained, which makes averaging of individual gains difficult. Also, the interpretation is questionable when relating a loss to the maximum possible gain. An absolute 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 ANG. Marx and Cummings have pointed out these and other shortcomings of $\langle g\rangle$ and proposed a different definition for negative values [11] which is presented in the following section.

2.2 Normalised Change

Since NIPPs do not necessarily have the same level of difficulty, negative gains are more likely to occur in the case of a test combination with a more difficult post-test. For our data from the FCI/SCI-combination, all $\langle g\rangle$ were negative. Therefore, this data cannot be plotted in *Fig. 1*. To make the ANG applicable to

[4] While this is especially true for class averages, the chances of obtaining negative gains are higher if the measure is applied to individual scores.

negative gains in an easy-to-interpret manner, the normalised change (NC) described by Marx and Cummings in [11] can be applied. While a positive absolute gain $\%\langle G\rangle$ is related to the maximum possible gain as in *Eq.* (1), a negative absolute gain $\%\langle G\rangle$ is related to the maximum possible loss, i.e.:

$$\langle c\rangle = \begin{cases} \frac{\%\langle S_f\rangle - \%\langle S_i\rangle}{100 - \%\langle S_i\rangle} & \text{if } \%\langle S_f\rangle \geq \%\langle S_i\rangle \\ \frac{\%\langle S_f\rangle - \%\langle S_i\rangle}{\%\langle S_i\rangle} & \text{if } \%\langle S_f\rangle < \%\langle S_i\rangle \end{cases} \tag{2}$$

Note that *Eq.* (2) represents the normalised change of *average* scores, and thus differs from the definition given in [11] Marx and Cummings discuss why taking the average of individual normalised changes is more appropriate than calculating the normalised change of average scores, but they also state that "for large numbers of students the numerical difference is small" [11]. For comparative purposes, we will hence pursue the averaging procedure in *Eq.* (2), which was also applied by Hake. The respective diagram to *Fig. 1* for *Eq.* (2) is shown in *Fig. 2* including data from the FCI/SCI test-combination. Each datapoint displayed corresponds to one of the eight years of the statics course described in *Section 1*. It can be seen that the IE-courses yield a greater $\langle c\rangle$ than the T-courses with $\langle c\rangle_T = -0.30$ and $\langle c\rangle_{IE} = -0.06$. This corresponds to the results found by [2].

Even though this definition allows a reasonable interpretation of negative gains and comparisons between gains, these are only valid for the same pre-test/post-test combination, i.e. a gain of FCI/SCI data can be compared only to FCI/SCI data. Therefore, the data shown in *Fig. 1* is explicitly not displayed here together with the FCI/SCI data as this would suggest that a comparison between these datasets was valid. This is due to the fact that a gain of zero can not be interpreted as "nothing was learned", as the pre- and post-test had a different level of difficulty.

However, it is feasible to compare the *differences* in average gains $\Delta\langle\langle c\rangle\rangle$ for any two courses, i.e. for IE- and T-courses:

$$\Delta\langle\langle c\rangle\rangle_{IE-T} = \langle\langle c\rangle\rangle_{IE} - \langle\langle c\rangle\rangle_{T}. \tag{3}$$

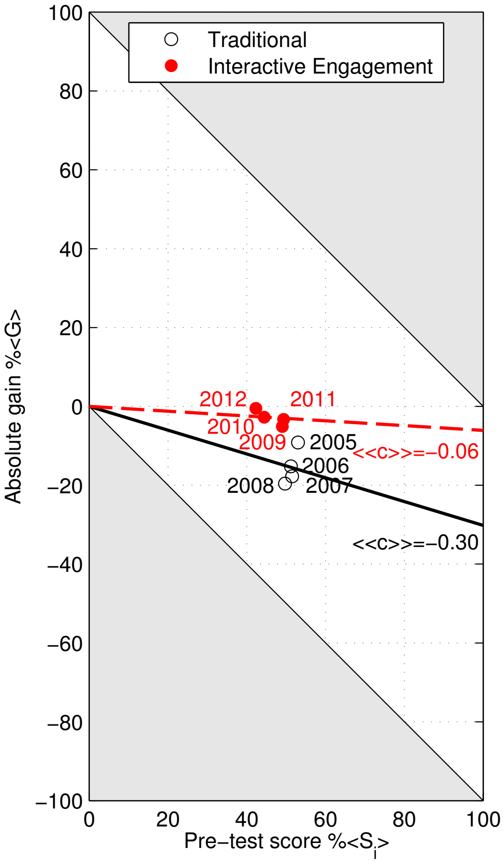


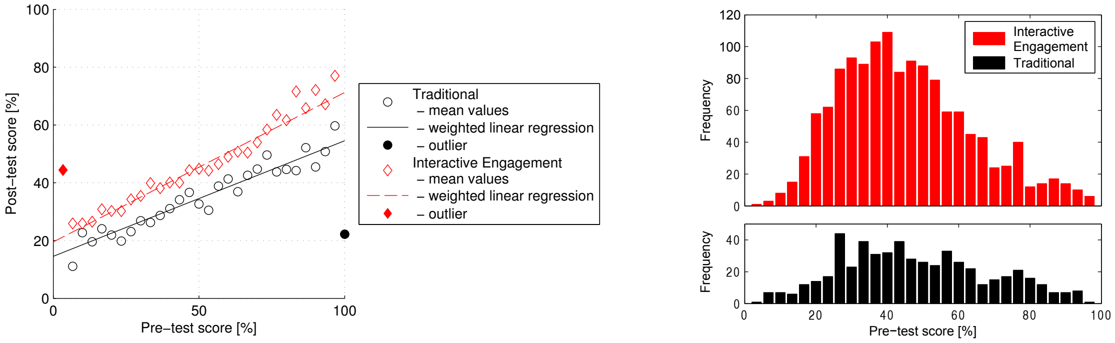
Fig. 2: Illustration of the normalised change, showing our data from the FCI/SCI test combination.

It is noteworthy that if *Eq.* (3) is applied to the data displayed in *Fig. 1* (Hake's data from IPPs) and *Fig. 2* (our data from NIPPs), respectively, these two values do not substantially differ from each other. For the data from *Fig. 1* we get $\Delta\langle\langle c\rangle\rangle_{IE-T} = 0.48 - 0.23 = 0.25$ while for the data from *Fig. 2*, we get $\Delta\langle\langle c\rangle\rangle_{IE-T} = -0.06 - (-0.30) = 0.24$. Assuming that IE-instruction does result in a greater learning effect and that the employed NIPPs combination does measure this effect, this supports the hypothesis that the NC is a valid statistical method for assessing NIPP data.

2.3 Weighted linear regression index

We propose an alternative measure for teaching effectiveness applicable to NIPPs, and refer to it as the weighted linear regression (WLR) index. For each student taking both, pre- and post-test, there exists a pair of pre- and post-test scores $(\%S_i, \%S_f)$. For each set of pairs with the same pre-test score, the mean value of the respective post-test scores $\%\langle S_f\rangle|_{S_i}$ are calculated. *Fig. 3(a)* shows these values for the data from the T- and IE-courses, respectively. We found that they can be well approximated by a linear model $y\left(x\right)$ with parameters w and m, where w is the value at $x = 50$, and m is the slope:

$$y\left(x\right) = w + m\left(x - 50\right). \tag{4}$$



(a) Weighted linear regression of post-test score in dependence of pre-test score.

(b) Histogram of pre-test scores.

Fig. 3: Average post-test score in dependence of pre-test score and frequencies of pre-test scores for T-courses and IE-courses.

	T	IE
w	34.567	45.390
m	0.400	0.517
R^2 of means	0.924	0.964

Table 2: Parameters of the weighted linear regression for T- and IE-courses

To calculate these parameters of the linear equation, a least squares approach was applied to the set of all individual post-test scores instead of on the mean values shown in the graph. Thus, the result is a *weighted* linear regression of the mean values shown in *Fig. 3(a)* as the number of occurrences for each pre-test score shown in *Fig. 3(b)* has been taken into account. Consequently, outliers of the average post-test scores as indicated in *Fig. 3(a)* do not have much influence. The resulting parameters of the model can be found in *Table 2*.

The representation of the WLR in *Fig. 3(a)* contains three main aspects of information: (a) we interpret a high degree of linearity as evidence for constant sensitivity of the tests across all levels of student ability, as the difference in mean post-test scores for a fixed difference in pre-test scores is the same anywhere on the pre-test range if the slope is constant. (b) The general post-instruction level can be deducted from the average value of the regression line, which is equal to w. (c) The discriminating effect of the instruction is represented by the slope m. It shows how subgroups of students with different pre-instruction levels responded to the instruction. A large positive slope would show that stronger students benefit a lot more from instruction than weaker students, while a large negative slope indicate that either the tests might not be valid in a way that they do not measure the intended criteria, or the instruction created misconceptions which rather affect students with good conceptual thinking. A slope close to zero represents an equalizing effect of instruction.

In order to estimate the goodness of fit of the linear model to the mean values, the coefficient of determination $R^2 = 1 - SS_{res}/SS_{tot}$ is calculated, with SS_{res} and SS_{tot} being the residual sum of squares and the total sum of squares, respectively [12]. For the calculation of R^2 the outliers[5] indicated in *Fig. 3(a)* were neglected. The values of R^2_{IE} and R^2_T show that the linear models with the parameters given in *Table 2* predict 92.4 % and 96.4 % of the variance in the average of the post-test score $\%\langle S_f\rangle$ for IE and T, respectively. This supports the assumption of a linear relationship. They are not to be confused with R^2 of individuals, which is only around 0.2 to 0.3 due to the greater variance in the individual post-test scores. The plotted IE-values in the high pre-test score range may also suggest a quadratic model. The least squares fit yields an adjusted R^2 of means of 0.974. While being more complex, the quadratic model does not result in a substantially better fit than the linear model. Also, the frequencies in the high pre-test score range are rather low and the variance of the mean values increases here (see *Fig. 3*). For these reasons, the quadratic approach is not considered any further.

[5] These data points are located at the extreme ends of the pre-test score range for which the frequencies are very low. Consequently, the resulting average values are not very reliable which additionally justifies the declaration as outliers.

		ANG	NC	WLR
i)	applicable to NIPPs	-	+	+
ii)	allow comparison	+	-	-
iii)	neutral element	+	-	-
iv)	linear	+[a]	+[b]	+
v)	bounded	+[a]	+	+

Table 3: Overview of fulfilment of criteria from Section 2 for each assessment method.

The multidimensional character of the WLR makes it more complicated to compare two courses, especially, if the regression curves in question intersect on the interval of $[0, 100]$. On the other hand, it supplies more detailed information on a course or differences between courses with respect to the three aspects mentioned above. A suitable way to reduce this measure to a one-dimensional index, and therefore provide easy comparison, if required, is to integrate *Eq.* (4) within the bounds $[0, 100]$, i.e calculate the area under the curve, which is equivalent to the parameter w introduced above.

From the data shown in the histogram in *Fig. 3(b)* it is evident that the average pre-test scores do not necessarily coincide with $S_i = 50\,\%$. Instead, they tend to lie below this mark. Therefore, one could argue that instead of looking at the value at $S_i = 50\,\%$ one should choose the value at the average, the median or the mode of the pre-test score. However, this would create a course-specific parameter and reduce comparability between courses. Also, more importantly, courses with a low average pre-test score but high average post-test score could be rated less successful than courses with high average pre-test score and low average post-test score, which is obviously wrong.

Comparing the WLR indices w of our data in *Fig. 3(a)* given in *Table 2* yields $\Delta w_{IE-T} = 10.8$. So, the IE-courses scored 10.8 percentage points higher on the post-test compared to the T-courses with respect to the deliberately chosen reference at the pre-test score $S_i = 50\,\%$. Since the correlation between pre-test score and average post-test score can be approximated to be linear, this value is an indicator for the performance of the entire course population, regardless of their pre-test scores.

3 DISCUSSION AND CONCLUSION

In this paper, we have presented three methods to evaluate combined data from pre- and post-tests. Of these three methods, two (ANG and NC) have been proposed by others and are already partially established as statistical methods for pre- and post-test data evaluation. The remaining method (WLR) has been proposed by us in order to account for data from non-identical pre- and post-tests. *Table 3* shows an overview of these methods and their assessment with respect to the criteria listed at the beginning of *Section 2*. The methods are each examined with respect to the criteria based on their intended employment on IPPs and NIPPs, respectively.

Only the NC and the WLR are applicable to NIPPs. The ANG does not consider negative gains, which can easily result from different levels of difficulty in NIPPs. If identical pre- and post-tests are used, the ANG does allow the comparison of results using different tests. In this limited realm, this is also true for both other measures. Neither of those, however, allows the comparison of data from different pairs of tests (NIPPs), due to possible non-identical levels of difficulty for such tests. Similarly, the different levels of difficulty of pre- and post-tests applied prevent the existence of a global neutral element to indicate “no learning”.

Comparing NC and WLR, we can see that, based on the criteria, they are equal. One difference that is not evident from *Table 3* is the occurrence of negative values (of NC) that do not necessarily imply a decrease in understanding. Similarly, one disadvantage of NC is that the zero might be easily misinterpreted as the value for “no learning”, which is only true for the special case of IPPs, whereas the unit line in WLR is not as easily recognised as such and therefore not as easily misinterpreted. On the other

[a] only for positive gains

[b] Due to the discontinuous definition around zero, in a strict sense, the NC is only linear if one considers only negative or only positive gains.

hand, a disadvantage of WLR lies in the prerequisite assumption that the linear model is a sufficiently accurate description of the data. As, on the one hand, this linearity is an interesting observation, and on the other hand necessary for the WLR, it should be investigated if such a linear behaviour can also be found with other courses and/or other test combinations. It is questionable, whether the WLR is a good measure if this linearity cannot be found with other courses and tests. But, since the degree of linearity can be related to the constant sensitivity of the tests, and our data show this linear tendency also for all courses from each individual year, it is reasonable to assume at this stage that it is a systematic and not a random phenomenon. Furthermore, as long as the degree of linearity of average post-test scores with respect to pre-test scores is sufficient, the WLR can be reduced to the one-dimensional index and thereby provide a good measure for teaching effectiveness. Under these circumstances, the slope of the regression line yields additional information about the relative effectiveness of instruction for different levels of student ability. Finally, even if cases were found where the degree of linearity was weak, a lot of information could still be drawn from looking at the results in the same way as in *Fig. 3(a)*. Based on these considerations, we think of the WLR as the superior way of comparing data from NIPPs.

Acknowledgements: We would like to thank Pieter Coppens for a number of helpful comments, as well as the *Institute of Mechanics and Ocean Engineering* at Hamburg University of Technology for allowing us access to their students.

References

[1] Steif, P. S. and Hansen, M. (2006) Comparisons between performances in a statics concept inventory and course examinations. *International Journal of Engineering Education*, **22**, 1070.

[2] Hake, R. 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)**.

[3] Dellwo, D. R. (2010) Course assessment using multi-stage pre/post testing and the components of normalized change. *Journal of the Scholarship of Teaching & Learning*, **10**.

[4] Novak, G. M., Gavrin, A., Patterson, E., and Christian, W. (1999) *Just-in-time teaching: blending active learning with web technology*. Prentice Hall series in educational innovation, Prentice Hall.

[5] McDermott, L. C. and Shaffer, P. S. (2012) *Tutorials in Introductory Physics*. Prentice Hall, updated preliminary 2nd edn.

[6] Brose, A. and Kautz, C. (2011) Identifying and addressing student difficulties in engineering statics. *Proceedings of the 2011 ASEE Annual Conference and Exposition*.

[7] Steif, P. S. and Dantzler, J. A. (2005) A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education*, **94**, 363–371.

[8] Hestenes, D., Wells, M., and Swackhamer, G. (1992) Force Concept Inventory. *The Physics Teacher*, **30**, 141–158.

[9] Huffman, D. and Heller, P. (1995) What does the Force Concept Inventory actually measure? *The Physics Teacher*, **33**, 138–143.

[10] Hestenes, D. and Halloun, I. (1995) Interpreting the Force Concept Inventory - a response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, **33**, 502–506.

[11] Marx, J. D. and Cummings, K. (2007) Normalized change. *American Journal of Physics*, **75**, 87–91.

[12] Neter, J., Wasserman, W., and Kutner, M. H. (1985) *Applied linear statistical models*. Irwin, second edn.