Active-learning settings and physics lectures: a performance analysis

G Schiltz¹, G Feldman² and A Vaterlaus¹

¹ ETH Zürich, Department of Physics, Zürich, Switzerland
² George Washington University, Department of Physics, Washington DC, USA

Abstract. We have divided a non-physics undergraduate student cohort into two parallel teaching settings. We offered a highly interactive flipped class (SCALE-UP) to one group of 52 students and a reformed LECTURE to the remaining 318 students. We compared student performance in both settings, based on a common midterm exam. The SCALE-UP students performed significantly better on conceptual problems, but for numerical problems, SCALE-UP and LECTURE students showed similar results.

1. Introduction
In 1993, Lillian McDermott advocated to actively engage students in introductory physics lectures as opposed to the passive listening in traditional lectures [1]. At that time, McDermott was at the forefront of a reformed teaching movement that, based on a still growing body of research evidence, thoroughly changed the way of teaching physics. Since then, a multitude of new teaching approaches have been developed, such as Peer Instruction [2], Problem-Based Learning [3] and SCALE-UP [4], just to name a few.

Whereas active learning now has become common in many U.S. universities, European universities still struggle in transforming their course formats. At ETH Zürich, we are offering introductory physics courses for almost all study programs (approximately 3000 students per year) and all of these courses are conducted as lectures. Since 2010, however, teachers from the ETH physics department have been supplementing their lectures with active-learning elements, mainly Peer Instruction, at very different levels. However, we still lack experience with collaborative small group settings, which inevitably would result in major resource investments (rooms, teachers, etc.). In spring 2017, one of us (G. Feldman), an expert teacher in SCALE-UP, spent his sabbatical at ETH Zürich where he introduced a collaborative group-learning environment within our local curriculum and applied to our specific European university culture.

During the spring teaching period, we have divided a non-physics student cohort into two parallel teaching settings, one focusing on skill development (SCALE-UP) and one focusing on content delivery (LECTURE). In this paper, we will compare the performance of students in both of these settings. Another GI Rep paper from this same conference [5] provides more detailed information on the implementation as well as on the students’ reactions.
The students in the SCALE-UP setting worked through different activities in small groups of 3-4 students each. Before each class, students started learning about a topic by doing assigned readings and online exercises via Mastering Physics [6]. In class, the student groups did activities that helped them understand the basic concepts from the home reading, and applied these concepts in hands-on experiments and problems. Lecturing was reduced to a bare minimum.

Apart from content delivery, the LECTURE setting included 40 demonstrations. A total of 37 conceptual clicker questions within a Peer Instruction environment engaged students interactively and provided immediate feedback to the instructor regarding their level of understanding.

318 students attended the LECTURE and 52 students were part of the SCALE-UP class. At the beginning, students were free to enroll for the SCALE-UP class, which was limited to 54 available seats. The teaching period extended over 13 weeks with three weekly contact hours. An additional weekly recitation session (1 hour) allowed students from both settings to discuss numerical problems in groups (15-25 students) together with teaching assistants.

2. Data considerations

To compare the learning achievements between the two settings, we collected performance data at three different intervals:

- Force Concept Inventory (FCI) [7] as a pretest in the first week: 30 multiple-choice questions, worth 1 point each;
- Midterm exam in week ten: 3 conceptual question sets and 3 numerical question sets, each worth 12 points;
- FCI as a posttest in the last week: 30 multiple-choice questions, worth 1 point each.

Participation in all of the three assessments was optional. Local ETH regulations did not allow us to make any of these assessments compulsory. Neither were we allowed to enforce class attendance. The only required element for students is a final exam, which will take place in January 2018. However, as an incentive, the result of the midterm exam can be counted for 10% of the final grade, but only if ameliorating the result of the final exam in January 2018.

With all three assessments being optional, the participation turned out to be extremely divergent. Out of the 52 SCALE-UP students, 44 took the pretest, 42 the miderm, and 37 the posttest. From the 318 LECTURE students, we had 153 participants for the pretest, 162 for the midterm, and 57 for the posttest. The pretest and the midterm results thus turn out to be the best available sample for comparing the performance of both settings.

![Figure 1. Mean scores of the FCI pretest. (a) covers the matching pairs (students who attended the FCI pretest as well as the midterm exam). (b) covers all available scores. Error bars correspond to the standard error.](image-url)
Since the students were not assigned randomly to the two settings SCALE-UP and LECTURE, it might be argued that only the best students would have joined SCALE-UP. If we consider the pretest to be a valid measure for prior knowledge, we can compare both groups depending on their pretest results (figure 1). The SCALE-UP group performed slightly better on the FCI pretest. However, this difference turns out to be statistically not significant. For all available pretest scores from figure 1b an independent t-test reveals $t(195)=1.67$, $p=0.097$ and $d=0.27$. Thus, we may assume that the difference of 1.6 points between LECTURE ($M=11.5$, $SD=5.4$) and SCALE-UP ($M=13.1$, $SD=6.3$) is within the limits of random variation.

Our further performance considerations will focus on the results of the midterm exam. For this reason, we have to consider the matching pairs, i.e. students who provided data for the pretest as well as for the midterm. This sample is limited to 35 students for the SCALE-UP group and to 92 for the LECTURE group. Comparing the FCI mean pretest scores from the matching pairs subset (figure 1a) shows similar results to the previous analysis of all available pretest scores. Again, the SCALE-UP group marginally outperforms the LECTURE students, but again this difference of 1.5 points between LECTURE ($M=12.2$, $SD=5.4$) and SCALE-UP ($M=13.7$, $SD=6.6$) turns out to be statistically not significant ($t(52)=1.18$, $p=0.243$ and $d=0.24$).

To sum up, based on the FCI pretest, students in the SCALE-UP group and in the LECTURE group had similar prior knowledge. Therefore, we may assume that there is no initial difference in students’ performance between the two groups. For our further considerations, however, we need to restrict the sample to those students who have participated in both assessments (matching pairs). Neglecting this limitation would potentially lead to invalid assumptions about the group’s initial performance.

3. Performance results

The midterm exam was given in week 10 and covered 6 sets of questions, each set worth 12 points for a total of 72 points. Related to the overall (complete) midterm results, the SCALE-UP group ($MD=41.7$, $SD=16.5$) outperformed the LECTURE group ($MD=34.2$, $SD=15.1$) by 7.5 points (figure 2a). This difference turns out to be statistically significant ($t(125)=2.44$, $p=0.016$ and $d=0.48$) and we can claim that in the midterm exam the SCALE-UP students performed better than the LECTURE students. Now we would like to quantify this gain and relate it to the different instruction methods for SCALE-UP and LECTURE.

![Figure 2. Mean scores of the midterm exam. (a) covers the matching pairs (students who attended the FCI pretest as well as the midterm exam). (b) covers all available scores. Error bars correspond to the standard error.](image)

For the performance analysis, we will consider the results from the FCI pretest as a baseline for prior knowledge and then calculate the gains based on the results of the midterm exam. An adequate gain calculation therefore is only possible within the matching pair sample, where we have student data from the pretest linked together with data from the midterm.
A common way to calculate and compare the gain of each group is to apply an analysis of covariance (ANCOVA) [8]. The major aim of ANCOVA is to estimate the learning gain achieved on the midterm that is not conditional on the pretest.

In a first step, we will consider the complete achievement in the midterm and then split the results according to the achievement obtained in the conceptual and numerical parts of the midterm.

3.1. Complete midterm performance

In figure 3 the scores of the midterm are plotted against the scores of the FCI pretest. For ease of comparison, we have unified the different scores of both assessments to a percentage scale. There is a correlation between the FCI pretest and the overall midterm results for the SCALE-UP group ($r=0.696$, $p<0.001$) as well as for the LECTURE group ($r=0.558$, $p<0.001$).

Students of each group who have achieved good results in the pretest are likely to achieve good results in the midterm. Furthermore, homogeneity of regression is not violated, which we tested statistically and also can be observed graphically by the almost parallel regression lines in figure 3. With those two preconditions being fulfilled for each data set, we are able to conduct the ANCOVA tests [9].

After controlling for the effect of the FCI pretest, ANCOVA results in a significant positive effect of the SCALE-UP group compared to the LECTURE group: $F(1,124)=4.26$, $p=0.041$ and an estimated gain of 7.10% (table 1, figure 4, complete data).

**Table 1.** Pairwise comparisons of the estimated mean midterm achievements (in %) resulting from ANCOVA after controlling for the effect of the FCI pretest. Gain is the difference between the estimated mean values of the SCALE-UP group and the LECTURE group: $G = M_{\text{SCALE-UP}} - M_{\text{LECTURE}}$.

| Midterm       | Gain (in %) | Std. error | Conf. interval     | p     |
|---------------|-------------|------------|--------------------|-------|
| complete      | 7.10        | 3.44       | [0.29, 13.90]      | .041  |
| conceptual    | 10.78       | 3.80       | [3.28, 18.32]      | .005  |
| numerical     | 3.31        | 4.14       | [-4.89, 11.51]     | .426  |
3.2. Performance gain in conceptual and numerical parts of the midterm

Three of the midterm question sets were conceptual and three had a major numerical component. The conceptual part included eight questions where students were asked to supply sketches or text together with justifications. The numerical part consisted of three common textbook problems with multiple subparts. Examples of conceptual and numerical questions are listed in the appendix.

Figure 4. Mean performance gain of the SCALE-UP group compared to the LECTURE group resulting from ANCOVA after controlling for the effect of the FCI pretest (data from table 1). Error bars correspond to the standard error.

We now can split the total midterm performance according to the results in the conceptual and in the numerical part (figure 5a). Figure 5b is added only for illustrative purposes and covers all available scores of the midterm. It shows no major differences compared to the matching pairs in figure 5a, even though the data sample for the lecture is considerably larger. We may thus conclude that the matching pair subsample reflects the overall sample relatively well.

Figure 5. Mean scores of the midterm exam split into the results linked to conceptual and to numerical questions. (a) covers the matching pairs (students who attended the FCI pretest as well as the midterm exam). (b) covers all available scores. Error bars correspond to the standard error.

In both parts of the midterm, the SCALE-UP group performed better than the LECTURE group. Independent t-tests show that the difference of 5.1 points on the conceptual part is highly statistically significant (t(125)=3.10, p=0.002 and d=0.60). However, the difference of 2.4 points on the numerical part turns out to be statistically not significant (t(125)=1.38, p=0.171 and d=0.27). The ANCOVA, while controlling for the effect of prior knowledge, comes up with similar results (table 1, figure 4). The estimated gain of 10.78% in the conceptual part is an important result (F(1,124)=8.07, p=0.005), but the gain of 3.31% in the numerical part, again, is statistically not significant (F(1,124)=0.64, p=0.426).
Thus, we can conclude that students taught under SCALE-UP conditions had a better conceptual understanding in the range of 11% as compared to those students taught in the LECTURE. Concerning the numerical skills, students in both groups achieved similar results.

4. FCI pretest/posttest results
As mentioned earlier, the student participation in the FCI posttest was small, especially if we consider the matching pair sample. Only 30 students from the SCALE-UP group and 38 students from the LECTURE group took part in all tests. Taking the FCI to be a valid instrument to measure conceptual achievement, figure 6 shows the normalized gains [10], which mirror our previous findings. The SCALE-UP group performed better than the LECTURE group in conceptual understanding and the gain difference is within the scope of our previous results.

Apart from the limited data points of our sample, the topical coverage of the FCI is restricted to kinematics and forces, which were only a fraction of the topics covered in our syllabus. For this reason, we show the FCI pretest/posttest results (figure 6) here only for illustrative purposes and will not include them in our discussion. Feldman et al. [5] provide a more detailed analysis of these results.

![Figure 6. Normalized gain of the FCI pretest/posttest results. Error bars correspond to the standard error.](image)

5. Discussion and conclusion
We have conducted a study on the performance gain in conceptual and numerical problems within different learning settings. We compared the learning gains of students taught in an active-learning SCALE-UP environment to those taught in a reformed lecture. The achievements in a common midterm exam revealed that the conceptual gains are stronger for students in the active-learning setting, but that the problem-solving performance of both groups is similar. Within the large body of physics research, equivalent performance studies were mostly carried out in the last decades of the 20th century and they compared traditional direct instruction to reformed lectures [11]. Recent studies on numerical performance gains that compare lectures to highly interactive classes such as SCALE-UP are sparse. Hoellwarth et al [12], while confirming our findings, report on much stronger conceptual gains for the active-learning group. This may be due to the fact that their contrasted lectures did not include any interactive elements such as Peer Instruction. McDaniel et al [13] report greater gains in conceptual understanding for the interactive-engaged groups, but lower performance results in numerical problem solving as compared to their traditional lecture groups. In our study, students from both teaching settings had access to common numerical problems that were discussed in the weekly recitation sessions. This consistent training may have enabled our SCALE-UP students to perform well in numerical problems. We can conclude that SCALE-UP, while increasing the conceptual understanding, does not compromise the skills of solving numerical problems, under the premise that those skills are trained explicitly.
The added conceptual gain of roughly 11% that we recorded for our SCALE-UP group was lower than expected. For instance, Hoellwarth et al [12] report on added conceptual gains of 200%. However, previous studies most exclusively make use of the FCI to measure the conceptual gain. Our calculation, however, relies on the eight conceptual questions from the midterm exam, which makes it difficult to compare the results.

We are aware that our study has some limitations. It is based on a single intervention with one cohort of students. In the near future, we will not be able to repeat a similar study with different student cohorts. Due to lack of control, we could only rely on a subset of possible data. For instance, we know nothing about the achievement of the 87 LECTURE students who have not participated in any of the assessments. However, even with those limitations, we believe that our study and the consequent findings provide a couple of important insights:

- We were able to compare the learning outcomes of a highly interactive setting to those of a reformed lecture. To our knowledge, this is the first study to contrast a modern (reformed) lecture with a collaborative interactive small-group setting such as SCALE-UP or Studio Physics.
- We carried out the study within our local curricular, time, legal and resource constraints. Those constraints cover the typical educational setting of a European research-intensive university. This is important because the majority of previous research is related to an Anglo-American context, facing rather different constraints. As an example, admission at ETH is only marginally restricted. Thus, selectivity in the undergraduate programs has a higher priority than retention.

Because everything was optional to the students, our data are not biased by explicit exam preparation (cramming). This means that we were able to measure and to compare the immediate learning effects in both settings. The current paper still lacks the results of the final exam and the feedback on students’ learning experience. Together with those data, being available in 2018, we will be able to supplement our study with assumptions on longitudinal learning effects and on students’ engagement.

Appendix
Example questions from the midterm exam (translated from German).

**Conceptual question**
An object explodes in a rigid pipe and breaks into 3 pieces. A 6 gram piece comes out the right end of the pipe, and a 4 gram piece comes out the left end at twice the speed of the 6 gram piece. From which end of the pipe does the 3rd piece emerge? Be sure to provide a careful explanation of your answers in words and/or equations.

**Numerical question**
A 3 kg block sits on top of a frictionless table, and a 2 kg block hangs over the edge of the table, connected by a pulley to the upper block. The hanging block is initially 2 m above the floor and is released from rest. (a) Using energy considerations only, find the speed of the 2 kg block just before it hits the floor. (b) Using work and energy considerations only, find the tension in the string.

**References**
[1] McDermott L C 1993 Guest Comment: How we teach and how students learn - A mismatch? *Am. J. Phys.* 61 295–8, DOI: 10.1119/1.17258
[2] Crouch C H and Mazur E 2001 Peer Instruction: Ten years of experience and results *Am. J. Phys.* 69(9) 970–7, DOI: 10.1119/1.1374249
[3] Duch B J 1996 Problem-based learning in physics: The power of students teaching students *J. Coll. Sci. Teach.* 15 326–29
[4] Beichner R J, Saul J M, Abbott D S et al. 2007 PER-Based Reform in University Physics vol 1, ed E F Redish and P J Cooney (College Park, MD) p 1–42 online available from http://www.per-central.org/document/serveFile.cfm?ID=4517 (accessed 12.10.2017)

[5] Feldman G, Schiltz G and Vaterlaus A 2017 Collaborative Group Learning in a Swiss Introductory Physics Class Proc. GIREP-ICPE-EPEC Conf. (Dublin) Journal of Physics: Conf. Series (in press)

[6] https://www.pearsonmylabandmastering.com/ (accessed 12.10.2017)

[7] Hestenes D, Wells M and Swackhamer G 1992 Force Concept Inventory Phys. Teach. 30 141–58. DOI:10.1119/1.2343497

[8] Weber E 2009 Quantifying Student Learning: How to Analyze Assessment Data Bull. Ecol. Soc. Am. 90 501–11. DOI: 10.1890/0012-9623-90.4.501

[9] Pituch K and Stevens J 2016 Applied Multivariate Statistics for the Social Sciences (New York: Routledge)

[10] Hake R R 1998 Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses Am. J. Phys. 66 64–74. DOI: 10.1119/1.18809

[11] Fraser J M, Timan A L, Miller K, Dowd J E, Tucker L and Mazur E 2014 Teaching and physics education research: bridging the gap Rep. Prog. Phys. 77 032401. DOI: 10.1088/0034-4885/77/3/032401

[12] Hoellwarth C, Moelter M J and Knight R D 2005 A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms Am. J. Phys. 73 459–62. DOI: 10.1119/1.1862633

[13] McDaniel M A, Stoens S M, Frey R F, Markow Z E, Hynes K M, Zhao J and Cahill M J 2016 Dissociative conceptual and quantitative problem solving outcomes across interactive engagement and traditional format introductory physics Phys. Rev. Phys. Educ. Res. 12 020141. DOI: 10.1103/PhysRevPhysEducRes.12.020141