Does it stick? - Investigating long-term retention of conceptual knowledge in mechanics instruction

Julie Direnga, Hamburg University of Technology

Julie Direnga studied General Engineering Science at Hamburg University of Technology in Hamburg, Germany from 2006 to 2010. Specializing in the field of mechatronics, she received a M.Sc. degree in 2014. Since March 2014, she is pursuing her Ph.D. in Engineering Education Research at the same institution.

Mr. Bradley Presentati, Hamburg University of Technology

Bradley Presentati completed a B.A. in English literature with an emphasis on creative writing in 2006 at UCSC. He is currently studying in the General Engineering Science bachelor program at Hamburg University of Technology with an emphasis on electrical engineering.

Mr. Dion Timmermann, Hamburg University of Technology

Dion Timmermann studied electrical engineering at Hamburg University of Technology, Hamburg, Germany. In his master thesis he worked on simulation methods for the signal and power analysis of high speed data links. He currently pursues his Ph.D. in the Engineering Education Research Group at Hamburg University of Technology, where he investigates students understanding in introductory electrical engineering.

Dr. Andrea Brose, Hamburg University of Technology

1998 Ph.D. in Mathematics, University of Colorado Boulder 1999-2008 Assistant Professor, lecturer and academic administrator, Department of Mathematics, UCLA 2008-2011 Engineering Education, Hamburg University of Technology Since 2011 Scientific Staff at Center for Teaching and Learning Since 2013 Head of Center for Teaching and Learning, Hamburg University of Technology

Prof. Christian H Kautz, Hamburg University of Technology

Christian H. Kautz has a Diplom degree in Physics from University of Hamburg and a Ph.D. in Physics (for work in Physics Education Research) from the University of Washington. Currently, he leads the Engineering Education Research Group at Hamburg University of Technology.

Does it stick? - Investigating long-term retention of conceptual knowledge in mechanics instruction

Abstract

By administering the Concept Assessment Tool for Statics (CATS) as a retest to engineering students and graduates, the retention of basic concepts in mechanical engineering is explored. Results from the retest from a sample population are compared to results from a larger population of posttests, i. e., tests after all relevant instruction. The sample is a subset of the population for which each member's posttest result is known. The sample population of 268 individuals is analyzed and grouped into three sub-populations based on the time interval between the posttest and retest. Overall we find that normalized differences between the sample retest and posttest is positive, showing a gain of understanding since the posttest at all retention intervals. It is hypothesized that gains in the retest relative to the posttest are a structural artifact of posttest being administered at the end of the course, but well before the exam period, an interval of sometimes up to eight weeks that is usually accompanied by intensive preparation and review of the course material. There is some evidence of forgetting when retention intervals are compared with one another. Sample members who remain actively engaged in the subject matter (as revealed by survey questions administered with the posttest) actually show learning gains and not only retention as a function of the retention interval.

1 Introduction

The purpose of instruction in engineering is to help prepare students for subsequent courses and future jobs. It might be an obvious statement, that the knowledge gained in engineering courses is meant to be retained, but is it really? While there is a large body of research that focuses on teaching effectiveness by measuring how much knowledge was gained during instruction, the question of how much of this knowledge is actually retained in the years following completion of the course is less frequently addressed, although it is equally important.

In this paper, we investigate the long-term retention of conceptual understanding in statics based on data from the Concept Assessment Tool for Statics (CATS).[1] Teaching effectiveness is often measured with such concept inventories by administering a pretest at the beginning of instruction and a posttest at the end of instruction. This usually takes place at the first and last lecture in the respective course. To better understand retention in this subject, we invited students to take the CATS again as a retest. The invitation to participate was open to all students and past students

who had been administered a posttest within the past ten years. We then compared students' posttest scores with their retest scores as a function of the retention interval (RI), the time interval between the post- and retest. This allowed us to investigate the influence of time on the retention of knowledge in mechanics. A survey was also administered to retest participants to gather data about other factors which might influence retention, i. e., the intensity with which they used the concepts being tested, for instance in their capacity as a teaching assistant (TA). We compare our findings with literature describing similar studies in other disciplines.

There is a large amount of research on knowledge retention in the fields of psychology and medical education. Custers et al.[2] published a review article on knowledge retention studies of three different types and from various disciplines, with an additional focus on basic science knowledge in the medical domain. The three study types consist of so-called (1) laboratory studies with short RIs of only a few hours or days, (2) classroom studies (like ours) with RIs of a few years, and (3) naturalistic studies with RIs spanning tens of years. In many naturalistic studies the knowledge is measured at the end of the RI. Knowledge at the beginning of the RI can only be reconstructed by e. g. counting the cumber of courses taken on the subject and respective grades achieved. Across all study types and disciplines, many of the results were adequately described by the Ebbinghaus forgetting curve.[3] This curve models retention over time as a fast decay at the beginning, transitioning into an ever smaller rate of decay as time passes. The parameters of the curve are significantly influenced by whether the material to be remembered is meaningful, and whether it was used during the retention interval. However, it is practically impossible to control for non-use of the subject matter during the RI in classroom studies. In naturalistic studies with very long RIs, retention might be seen as reaching a point of saturation, indicating the existence of a permanently stored knowledge. In 2011, Custers et al. reported the results from their own study, which for the most part reproduced the findings of the other related studies[4]. Most of those studies, however, are on rote knowledge. Is it the case that these findings are relevant for conceptual knowledge, as it is often required in physics and engineering education?

Forgetting may depend on the subject matter and teaching methodology. Examining retention studies in physics instruction specifically, Francis et al.[5] reported in 1998 that they observed only a small decline in the Force Concept Inventory (FCI) scores of 127 students with RIs over one, two, and three years. They concluded that the knowledge is largely retained, and that "some forms of instruction (but not necessarily all) do achieve fundamental shifts in students' conceptual frameworks". Pollock[6] supported this claim in 2009. He used the Brief Electricity & Magnetism Assessment (BEMA)[7] to show that different pedagogies strongly influence retention of conceptual knowledge. Students from freshman courses, where *Tutorials in Introductory Physics* by McDermott and Shaffer[8] were used scored higher when tested after finishing upper-division physics courses, compared to students from the control group. His results indicated that upper-division courses do not further increase the score above that attained in the posttest administered in the freshman course.

Pawl et al.[9] conducted a study in 2012, investigating the analytical and conceptual freshman physics knowledge of 56 seniors with a fixed 4 year RI. The authors found that while there was a loss in analytical knowledge, which strongly depended on the intensity of use, the total scores on the conceptual Mechanics Baseline Test (MBT)[10] did not change significantly. Upon closer inspection, however, it was evident that knowledge was not simply retained. The test could be

broken into two distinct components: one where significant gains in knowledge were made, and the other where there were significant losses of what was gained in the course. The authors therefore stress that an analysis of changes should be carried out for each item or concept instead of the total test score.

2 Methodology

2.1 Institutional practice

At the Hamburg University of Technology (TUHH) we regularly assess teaching effectiveness in the introductory mechanics course, which covers statics, using pre- and posttests. At the end of this course students are given the CATS as a posttest. The CATS has been administered in every statics course since 2006. In the past, this continuous assessment was helpful in identifying frequently occurring conceptual difficulties, developing course material according to those difficulties, and assessing the success of this material in the classroom.[11,12] Starting in 2009 the teaching methods in the introductory mechanics course were changed gradually. Tutorial worksheets in the style of *Tutorials in Introductory Physics* by McDermott and Shaffer[8] were introduced and Elements of Just-in-Time Teaching (JiTT)[13] were added in 2011. Three different instructors taught this course over the last ten years.

2.2 The Concept Assessment Tool for Statics (CATS)

The CATS was developed as a formative assessment, diagnostic tool by Steif and Dantzler[1]. As state of the art test in its domain, it is widely used in statics education. The test consist of multiple choice questions (MCQs), with one correct answer and four distractors per question.

The CATS - formerly known as Statics Concept Inventory (SCI) - has 27 questions, which can be grouped into nine concepts with three questions each. Steif and Hansen[14] described the concepts as follows: drawing forces on separated bodies, Newton's 3rd law, static equivalence, roller joint, pin-in-slot joint, loads at surfaces with negligible friction, representing loads at connections, limits on friction force, and equilibrium.

The authors of the CATS do not suggest a time limit. Steif and Dantzler[1] initially gave the students an entire week in which to finish the test. Steif and Hansen[14] later reported that "[t]ime limits of 50 to 60 minutes were imposed for tests taken in class". At TUHH we imposed a time limit of 32 minutes for the pre- and posttest administered in the course respectively. We decided to use the same limit of 32 minutes for the retest. Additionally, before starting the test, the participants were asked to answer eight survey questions.

2.3 Survey questions

Along with the questions of the CATS we posed eight survey questions. One of these, a question relating to their experience as a TA in mechanics or mechanics related course, plays a role in our

subsequent analysis. All of the survey questions are listed in Appendix A. Beyond the survey questions we cannot control for non-use of the knowledge in questions within the scope of this study.

2.4 Administration of the retests

Although the posttest was administered exclusively in a paper-based format, we decided to offer the students to take the test in either paper-based format on campus or online. By offering the test online we intended to increase the response rate by reaching those students who have already left TUHH or who are currently not on campus. The primary reason for offering the test paper-based was to create greater visibility on campus. Students who saw the advertisements on campus could go directly to the testing room, while students who read the advertisement via email could immediately participate online.

In general, online and paper-based tests cannot be treated as the same tools of assessment, even if the test items are identical. There are certain situations where paper tests are easier for students, for example when long texts that require scrolling, or items including graphing or geometric manipulations are involved.[15] Unfamiliarity with the use of computers can also negatively influence test results for individuals who might have otherwise answered correctly in a paper-based test. Likewise, things that can easily be done on paper-based tests, e. g. switching from one item to another, backtracking, or comparing two or more items, must be handled differently in an online test. The additional complication that this introduces depends heavily on the implementation.

There are, however, certain conditions under which online tests can be seen as equivalent to paper-based tests.[16] The nature of our population is such, that offering an online test is not such an unreasonable proposition. Our potential participants are all “digital natives” at a university of technology. The items are multiple choice questions, and the test requires no calculations, graphing or geometrical manipulations. Additionally, we designed our online test and data logging such that we could guarantee easy switching between items, and monitor the screen size to issue a warning to the participant, or declare the data invalid if the test was done on a screen size which was too small to display all of a single item.

The possibility of cheating on an online test is always present. On our test it is reduced by the nature of the individual items, as these can hardly be answered by means of a web search. The correct answer requires conceptual understanding. Most importantly, test performance itself was neither rewarded nor punished, which eliminates a large part of the incentive to cheat. On the other hand, the lack of a performance-based reward motivates the question of whether participants took the online test seriously. A study by Germine et al.,[17] comparing supervised administration of a computer based set of tests in a lab setting with controlled participants against “uncompensated, anonymous, unsupervised, self-selected participants”, showed that the lack of supervision in the online test format does not reduce the quality of the data. They showed that the variability across samples is comparable, and that online participants took the test just as seriously. There was little evidence for cheating when they compared matched groups. Furthermore, the CATS was initially administered as a computer-based test outside of a controlled environment.[1] Since Steif and Hansen[18] report that no differences could be seen between the scores

of online and paper-based administration, we felt that it was safe to administer the tests in both modes. Finally, we clearly stated the goal of this study and emphasized that taking the test seriously is very important for us to receive high quality data. As will be discussed in section 2.6, we found no significant difference between the scores of participants who did the online test vs. the ones who did the paper-based test.

2.5 Incentives

We invited all students and former students to retake the posttest they were administered at the end of the static course they have taken. Each student's retest responses, score, and set of survey responses was matched to their results on the posttest. Because we were able to match each student to themselves, our study retains validity despite the fact that it might attract an unrepresentative sample of the population. The following goals influenced our study design: We wanted to obtain

- a large sample size with respect to the population, which we define as all students who were administered the CATS as a posttest at TUHH,
- samples from multiple RIs, and
- high quality data.

After the first semester, students at our institution do not follow a strict study plan. For this reason, an attempt to administer the retest in the context of a particular lecture, or even a set of various different lectures, was unlikely to yield satisfactory results. We therefore invested our resources in raising awareness about the study and relied on voluntary participation. The downside of this strategy is that it is prone to bias due to self-selection. In order to reduce this effect, we advertised the study via multiple channels and offered a variety of incentives.

Lotteries have been shown to generate a high rate of response despite a limited budget. They have been found more effective as an incentive if there are several smaller prizes associated with higher chances of winning instead of one single prize associated with a lower chance of winning.[19] For this reason, we decided to offer 15 small prizes to our participants via lottery. Participation in the lottery was voluntary. As an additional incentive, participants in the paper-based test were given candy bars upon completion of the retest as was advertised. This helped compensate for the fact that the retest was administered in a room that some may have perceived as being inconvenient to reach. All participants had the opportunity to receive their score in both the posttest and retest, allowing them to rate their performance and command of statics concepts.

In the following sections, we will present the results, paying particular attention to (1) the characteristics of the sample and (2) the average normalized change of test scores from posttest to retest with respect to RI. We expected to see an Ebbinghaus forgetting curve, or alternatively, no change on the overall average test score.

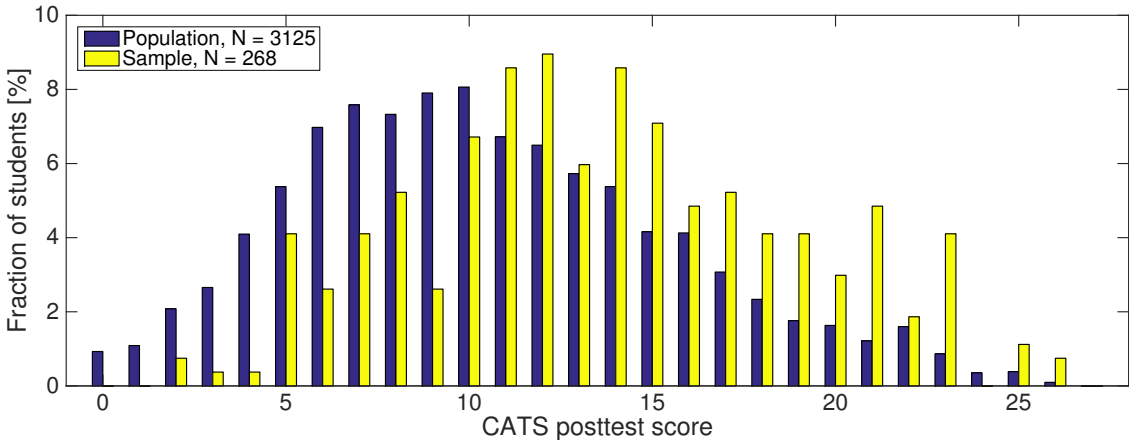


Figure 1: Influence of self-selection on the sample. Students with higher posttest scores were more likely to participate.

2.6 The sample

With the advertisement and incentives offered, we attracted a total of 301 participants for the CATS retest, of which 291 retest results were deemed usable, while 10 results where dismissed either because they spent too little time (less than 10 minutes as suggested by Steif and Hansen[20]) or it was not possible to match posttest and retest. approximately 4 % of the students from whom we have posttest data. Of these, 72 % (193) took the test online whereas 28 % (75) did the test in paper-mode. The retest scores do not significantly differ between online and paper modes (two-sample t-test, $p = 0.38$, $p < 0.05$), which confirms our previous assumption that we can treat them as equivalent. Results from online participants who answered less than one third of the test, or who submitted their answers in less than one third of the allowed time were considered "unserious" attempts, and the associated data was rejected from further analysis. In using these thresholds, we follow Steif and Hansen,[14] who also excluded data from students who took less than ten minutes to complete the test. Their justification for doing so was that the scores associated with these attempts were comparable to guessing. From our own experience with the paper-based tests we can say that these limits are conservative, but serve to exclude the most corruptive data. There were 23 such cases, resulting in a final sample size of 268.

We expected our sample to differ slightly from the population due to self-selection. In fact, the average posttest score of the sampled students was higher than that of the entire population (13.7 ± 0.3 out of 27 vs. 10.5 ± 0.1). This three point difference is statistically significant (one-way ANOVA, $F(1, 3391) = 93.5$, $p < 0.001$). At the same time, the variance in posttest scores is similar. The difference is simply an upward shift of the mean score (Figure 1). A factor that might contribute to this shift is that our sample likely does not include students who have dropped out in the intervening years. One would expect the scores for these students on the posttest to be in the lower ranges. Given the size of our sample it proved convenient to bin the data into three groups of RIs with approximately equal sample sizes - short (RI $\leq$ 2 years, $n = 99$), medium (2 $<$ RI $\leq$ 4 years, $n = 98$), and long (4 $<$ RI $\leq$ 9, $n = 71$).

As can be seen in Figure 2, students with a long RI are slightly under-represented, possibly be-

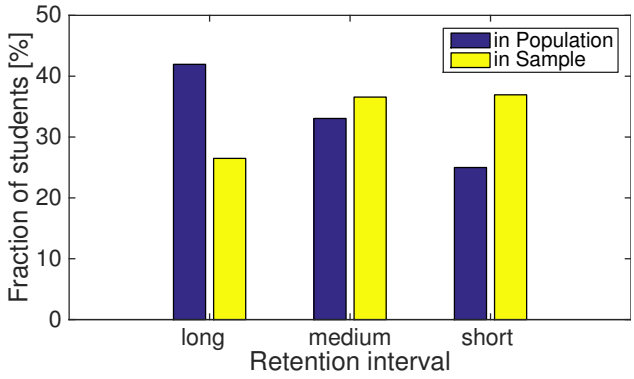


Figure 2: Differences between the population and the sample with respect to the RI groups (CATS only)

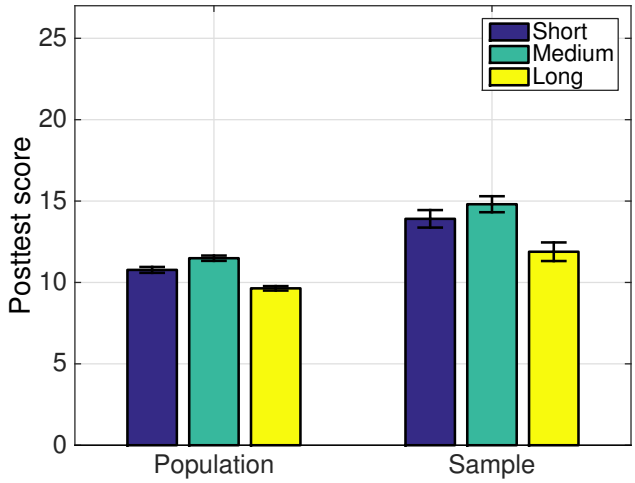


Figure 3: Comparing the post-scores of the CATS population and sample for all three RI bins. Error bars indicate the standard error of mean.

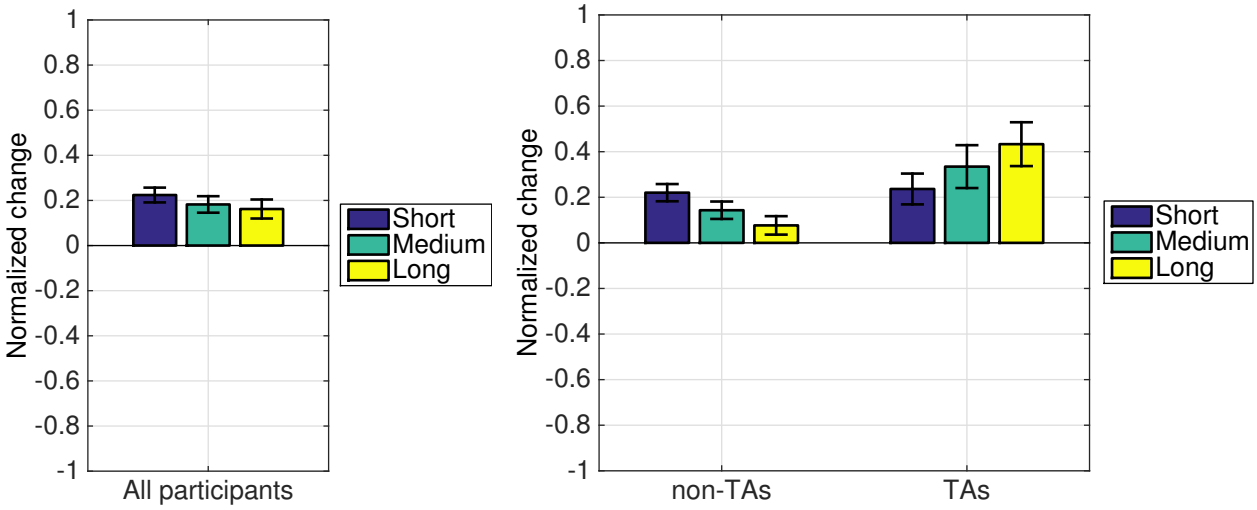
cause these students are not closely connected to the university anymore, and might not have been aware of the study. Students with medium RIs are represented at a rate commensurate with their presence in the population. Consequently, students with short RIs are over-represented.

Figure 3 shows the posttest scores of the population and the sample, each split into the three RI groups. Comparing each of the RI groups separately, we see that the posttest scores of the sample are always significantly higher than those of the population.

Looking only at the population, we see that there are significant differences among all three RI groups (10.8 ± 0.2 short, 11.5 ± 0.2 medium, and 9.6 ± 0.1 long). In our sample, only the long RI group had significantly lower posttest scores than the other two groups (11.9 ± 0.6 for long vs. 14.8 ± 0.5 for medium and 13.9 ± 0.5 for short, see Figure 3). As we do not have an equal baseline over all three RI groups, we performed the analysis by using the normalized change proposed by Marx and Cummings[21], that is the ratio of the gain to the maximum possible gain or the loss to the maximum possible loss (see Appendix B), instead of absolute score shifts from posttest to retest.

One of the survey question asked whether participants had experience working as a teaching assistant or tutored students in a mechanics, physics or design theory course (see Appendix A, question 4.). Almost 21 % of the participants stated that they were TA for mechanics courses. We estimate that about 10 % of our population is made up of former mechanics TAs. Thus, students who were TA in mechanics are strongly over-represented. This is, however, not a significant factor influencing the higher average posttest score in our sample. Although TAs have a mean score of 14.6 ± 0.7 points on the posttest and non-TAs only have a mean score of 13.4 ± 0.4, a two-sample t-test shows that the samples probably do not come from different populations.

3 Results



(a) On average, there is positive gain in all groups. No significant difference can be found between the groups.

(b) When mechanics TAs are excluded from the sample, the long RI group shows significantly less gain than the short RI group (left). Participants with TA experience show higher gain in the long RI group (right).

Figure 4: Normalized change of CATS scores from posttest to retest for all RI groups. Errors are standard error of the mean. The values can also be found in Table 1.

3.1 The influence of time on the retention of knowledge in our sample

Figure 4(a) shows the mean normalized change for the different RI groups for the entire sample. We detect no significant differences between the RI groups with respect to the average normalized change in test scores from posttest to retest. A one-way ANOVA yields $F(2, 267) = 0.72$, $p = 0.49 > 0.05$. Comparing the results to our previously stated expectations, we see that the Ebbinghaus forgetting curve[3] does not provide a suitable fit for our data, as there is no decrease in test scores over time. Our alternative prediction, that the average scores would not change from posttest to retest, also fails to accurately describe our data, as we can observe a gain significantly

different from zero ($p < 0.001$). About 63 % of the participants experienced a gain, 29 % experienced a loss, and 8 % remained by their original score. The predominance of those experiencing a gain might be due to the bias in our sample. It could also be due to the fact that the RI was not actually a period of non-use for most of our participants. Two cases of intensive knowledge use - TA activity and exam preparation - will be discussed in the following sections 3.2 and 4.1.

We also did an analysis on the concept level. Although we can see a gain in most concepts and a loss in others, we do not see a clear and consistent interpretation.

3.2 Accounting for use of knowledge and TA activity

On the survey questions, the participants reported how often they use mechanical loads concepts (see Appendix A question 5.). An ANCOVA analysis with the levels of knowledge use (often, seldom, and never) as covariates does not reveal a statistically significant difference in the normalized change of test scores from posttest to retest between any of levels of knowledge use. This is in line with Pollock's findings that follow-up courses do not increase the test score.[6] It is widely known that learning is strongly enhanced by giving instruction to a peer[22,23]. For this reason, we investigated the influence of experience as a TA in mechanics on retest performance.

As can be seen in Table 1, the 60 participants in our sample who reported to have been mechanics TAs are quite evenly distributed over the three RI bins. When we perform a separate analysis on the subgroups of TAs and non-TAs, we see two opposing trends. While the normalized change decreases over time for the non-TAs, the TAs show increased gain for longer RIs. This gain, however ist not significant (two-sample t-test, $p < 0.05$). We no longer see a significantly positive normalized gain for the non-TAs (one-sample t-test, $p > 0.05$). The TA group associated with the long RI, on the other hand, shows a greater positive normalized change of approximately 0.4. A normalized gain of this magnitude is denoted as a "medium" gain by Hake,[24] though this categorization was meant for cases where there is instruction.

The short RI group is comparable in both subgroups, while the normalized change in the medium group is slightly smaller for the non-TAs compared to the TAs, yet still significantly positive (one-sample t-test, $p \ll 0.001$).

An overview of the normalized change values is given in Table 1.

	short RI		medium RI		long RI	
	n	mean normalized change	n	mean normalized change	n	mean normalized change
all	99	0.22 ± 0.03	98	0.18 ± 0.04	71	0.16 ± 0.04
non-TAs	76	$\mathbf{0.22 \pm 0.04}$	78	0.14 ± 0.04	54	$\mathbf{0.08 \pm 0.04}$
TAs only	23	0.24 ± 0.07	20	0.33 ± 0.09	17	0.43 ± 0.10

Table 1: Average normalized changes of the RI groups. Errors are standard errors of means. Means of samples in bold are significantly different from each other ($p < 0.05$) within the table row.

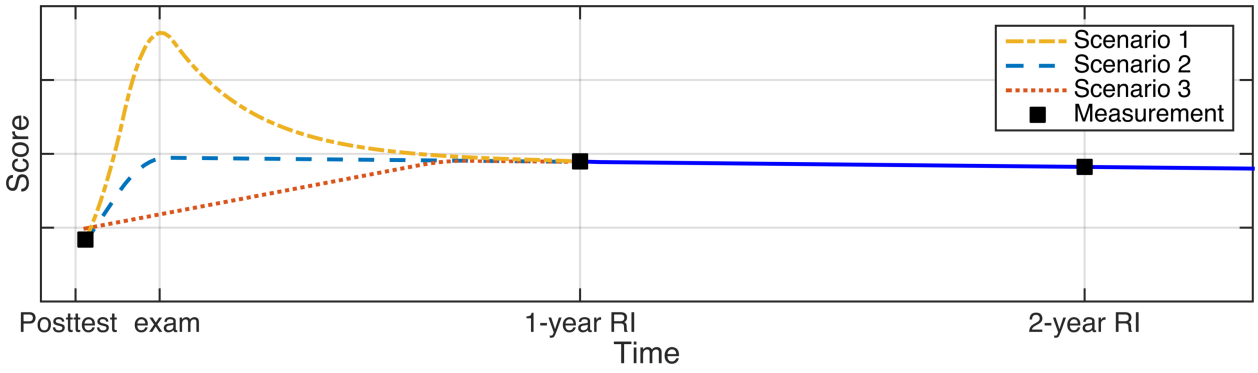


Figure 5: Schematic representation of possible score development over time for the period between posttest and exam.

4 Discussion and conclusion

4.1 The delayed-exam factor

None of the studies mentioned above found a gain in scores on a retest. The German higher education system has a feature less common (if at all) in other parts of the world, that might explain the observed positive gain from posttest to retest. Exams for classes are given over a period of two resp. three months after classes ended, depending on the semester being in the Winter or Summer. Few students study continuously throughout the time when classes are held, but start studying in earnest once classes are over. Because we administer the posttest in the last lecture held in the semester, a greater (positive) learning gain is captured less by our posttest and more so by the retest. Due to this tendency of our students to continue learning after the posttest, it is difficult to compare our results to those of other studies.

Figure 5 shows a schematic plot of possible continuous timelines of scores. The full lines indicate intervals for which we have a general idea about the trend of how conceptual knowledge develops qualitatively over time. From the data of this study we see a predominant additional gain after the posttest. At this point, the two groups discussed above tend to diverge. Important here is the time interval between the posttest and the first retest (one-year RI), which is still a mystery. We know that there must be a gain and therefore propose three plausible scenarios:

Scenario 1: The learning gain continues at a much faster progress in the exam studying period, reaching its peak at the time of the exam. Afterwards, forgetting happens, which might be described by something similar to the Ebbinghaus curve.[3]
Scenario 2: The learning gain continues until the exam is over, then remains at the final level. This seems plausible because most students quit studying immediately after the exam (see section 4.1).
Scenario 3: The learning gain continues, but the final level is not yet reached at the time of the exam. It takes some time to sink in, but the final level is reached within a one-year period.

In order to investigate the development between posttest and one-year retest, we are planning to introduce an early retest at the beginning of the second semester.

4.2 Conclusion

The fact that the mean retest scores of both TAs and non-TAs was larger than the respective posttest scores was unexpected – but certainly very favorable. The reasons for this unusual behavior are still unknown, but we suspect that the protracted exam study period after the posttest contributes to this effect. Further investigations into this effect are being conducted.

Comparing the groups with short, medium, and long RIs, there is a slight decline of the mean normalized change over time for the whole sample. This decline is not significant. When the sample is split in students that were a TA in mechanics and those who were not, different trends emerge. While the mean normalized change for the TAs increases from the short to the long RI, the mean normalized change for the non-TAs decreases. The increase of about 0.19 for the TAs, which is comparable to the normalized change observed during a tradition lecture,[24] is not significant. The decrease of the mean normalized change of the non-TAs from the short to the long RI, however, is significant.

Considering the main question of this paper – does conceptual knowledge in mechanics stick? – the normalized change of test scores seems to differ depending on the level of use, as indicated by the group of TAs. For the whole sample of students, however, the normalized change of the test score was not significantly influenced by the duration of the RI.

Appendix

A Survey Questions

Participants of the retest study where asked additional survey questions regarding:

1. if they had taken physics and/or math at an advanced or intermediate level or not at all during their last two years of high school,
2. their grade in the mechanics and/or physics course taken at TUHH*,
3. the number of attempts they required to pass the mechanics and/or physics exam*,
4. if they were a teaching assistant or tutored students in a mechanics, physics or design theory course,
5. how often they use topics from statics (mechanical loads) in their studies or in their job,
6. how long ago they were last using mechanical loads,
7. details about their current occupation,
8. academic degrees they had obtained.

B Normalized Change

Normalized Change The idea of using a normalized gain to evaluate the amount learnt in a course is probably widely known due to its employment in Hake's study on interactive-engagement vs. traditional teaching methods[24]. The normalized gain relates the absolute gain to the maximum possible gain and thus can be expressed as

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

where $\%\langle G \rangle$ is the absolute gain of class average from pretest to posttest scores, and $\%\langle S_i \rangle$ and $\%\langle S_f \rangle$ are the final (post) and initial (pre) class averages, respectively, all given as a percentage of the maximum possible score.

Equation 1, however, does not consider the possibility of losses. Relating an absolute loss to a maximum possible gain does not make much sense, although it is mathematically possible. While this is not a problem for the situation in Hake's study[24], we are likely to encounter losses from posttest to retest for individuals as well as for group average scores. We therefore use the normalized change proposed by Marx and Cummings[21], which relates losses to the maximum possible loss instead of the maximum possible gain:

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

References

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

[2] Custers, E. J. F. M. (2010) Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, **15**, 109–128.

[3] Ebbinghaus, H. (1885) *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

[4] Custers, E. J. and ten Cate, O. T. (2011) Very long-term retention of basic science knowledge in doctors after graduation. *Medical education*, **45**, 422–430.

[5] Francis, G. E., Adams, J. P., and Noonan, E. J. (1998) Do they stay fixed? *The Physics Teacher*, **36**, 488–490.

[6] Pollock, S. J. (2009) Longitudinal study of student conceptual understanding in electricity and magnetism. *Physical Review Special Topics-Physics Education Research*, **5**, 020110.

[7] Ding, L., Chabay, R., Sherwood, B., and Beichner, R. (2006) Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, **2**.

[8] McDermott, L. C., Shaffer, P. S., and Group, U. o. W. P. E. (1998) *Tutorials in introductory physics*. Prentice Hall.

[9] Pawl, A., Barrantes, A., Pritchard, D. E., and Mitchell, R. (2012) What do seniors remember from freshman physics? *Physical Review Special Topics - Physics Education Research*, **8**, 020118.

[10] Hestenes, D. and Wells, M. (1992) A mechanics baseline test. *The Physics Teacher*, **30**, 159–166.

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

[12] Direnga, J., Timmermann, D., Brose, A., and Kautz, C. (2014) A statistical method for assessing teaching effectiveness based on non-identical pre-and post-tests. *Proceedings of the SEFI 2014 Annual Conference*, Birmingham, UK, Sep.

[13] Novak, G., Gavrin, A., Christian, W., and Patterson, E. (1999) *Just-In-Time Teaching: Blending Active Learning with Web Technology*. Addison-Wesley.

[14] Steif, P. S. and Hansen, M. A. (2006) Feeding back results from a statics concept inventory to improve instruction. *Proceedings of the 2006 American Society of Engineering Education Conference and Exposition*.

[15] Keng, L., McClarty, K. L., and Davis, L. L. (2006) Item-level comparative analysis of online and paper administrations of the texas assessment of knowledge and skills. *Applied Measurement in Education*, **21**, 207–226.

[16] Karkee, T., Kim, D.-I., and Fatica, K. (2010) Comparability study of online and paper and pencil tests using modified internally and externally matched criteria.

[17] Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012) Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, **19**, 847–857.

[18] 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.

[19] Deutskens, E., De Ruyter, K., Wetzels, M., and Oosterveld, P. (2004) Response rate and response quality of internet-based surveys: an experimental study. *Marketing letters*, **15**, 21–36.

[20] New practices for administering and analyzing the results of concept inventories, volume = 96, number = 3, journal = Journal of Engineering Education, author = Steif, Paul S. and Hansen, Mary A., year = 2007, pages = 205–212, issn = 2168-9830, doi = 10.1002/j.2168-9830.2007.tb00930.x, url = http://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2007.tb00930.x/abstract.

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

[22] Crouch, C. H. and Mazur, E. (2001) Peer instruction: Ten years of experience and results. *American Journal of Physics*, **69**.

[23] Schmidt, B. (2011) Teaching engineering dynamics by use of peer instruction supported by an audience response system. *European Journal of Engineering Education*, **36**, 413–423.

[24] 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)**.