Further investigations into the effectiveness of collaborative group exams

Joss Ives

Dept. of Physics and Astronomy
University of British Columbia, Vancouver, Canada
joss@phas.ubc.ca
learnification.wordpress.com
@jossives

Scan for contact info

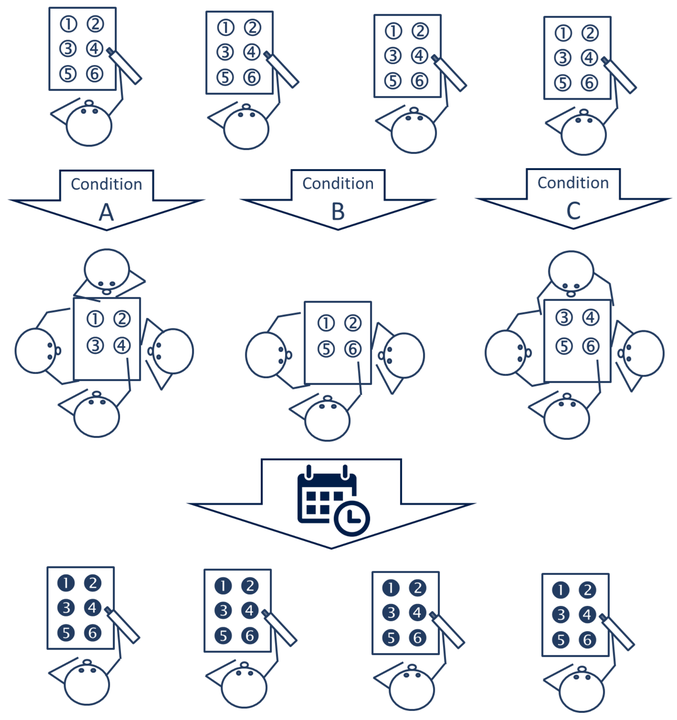
Summary

To quantify the learning impact of collaborative group exams, a randomized crossover design was used in 2014 and 2015 in an introductory calculus-based physics course where each student participated in both the treatment and control groups. Questions from each of the two midterms were designed to form near-transfer pairs with the end-of-course diagnostic, which was used as a retest to measure learning.

In both years, improved learning was shown in the treatment group for retest questions associated with the second midterm (retest within 2 weeks of the midterm). The 2014 data show no improved learning in the treatment group for retest questions associated with the first midterm (6-7 weeks prior to retest) and the 2015 data show a decrease in learning for retest questions associated with the first midterm (4-6 weeks prior to retest).

A likely explanation for this difference is that there is a time-based decay of the learning impact from the groups exams. However, additional studies are needed to investigate the difference in results between 2014 (null) and 2015 (decreased learning) for the longer times between midterm and retest.

Two-stage collaborative group exams and study design



All students first completed the midterm exam individually
- 2014: Midterm 1: n = 679, Midterm 2: n = 673
- 2015: Midterm 1: n = 701, Midterm 2: n = 703

Treatment: Immediately after the individual exams are collected, students self-organized into collaborative groups of 3 or 4 and retook a subset of the original exam questions (different subsets for conditions A-C).

Retest: The end-of-term diagnostic contained near-transfer questions that partnered with those from the original exam.

The time between the first midterm (questions 1.1-1.x) and the diagnostic was 6-7 weeks in 2014 and 4-6 weeks and the time between the second midterm (questions 2.1-2.x) and the diagnostic was 1-2 weeks in 2014 and 1 day to 2 weeks in 2015.

A mixed-effects logistic regression showed improved learning for short timeframes (up to 2 weeks) but null (2014) or decreased learning (2015) for longer timeframes (4-7 weeks)

The model:

In the following mixed-effects logistic regression model, a positive β_3 indicates the group exams had a positive effect on retest success. The analysis was run separately for each year and for the retest questions associated with midterm one (Q1.1-1.x) and for those associated with midterm two (Q2.1-2.x):

$$\text{Log\_odds}(\text{Retest\_success}_{ijk}) = \beta_0 + \beta_1 \times \text{Pre} + \beta_{2k} \times \text{Question}_k + \beta_3 \times \text{Treatment} + \varepsilon_i,$$

where,

- $\text{Retest\_success}_{ijk}$ is the (binary) success on the learning test of Student_i on Question_k;
- Pre is a binary variable that indicates if Student_i answered correctly the question isomorphic to retest Question_k.
- Question_k is a categorical variable representing question number and account for differences in question difficulty; and
- ε_i is a random intercept for Student_i which accounts for differences in student ability

Results:

Shorter timeframes (retest given within 2 weeks of the *2nd* midterm)

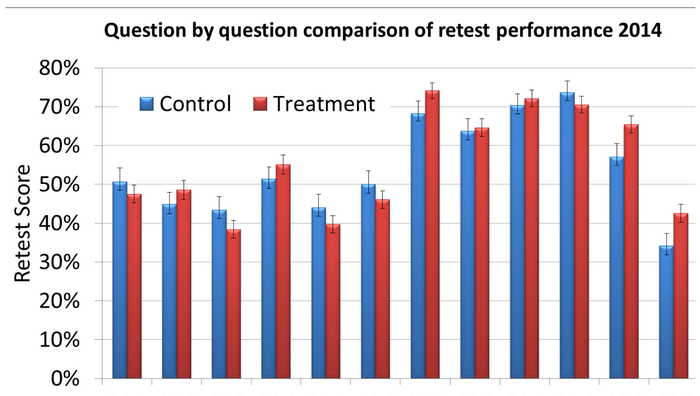
- Treatment (collaborative group exam) predicted success for retest questions Q2.1-Q2.x, (2014: β_3 = .203, SE = .079, p = .011 & 2015: β_3 = .363, SE = .083, p < .001)
- Expressed as odds ratios, the odds of answering a question correctly on the learning test versus not answering it correctly increased by a factor of 1.22 (95% CI [1.05, 1.43]) in 2014 and a factor of 1.44 (95% CI [1.22, 1.69]) in 2015 for those in the treatment as compared to the control.
- The fits between the model and data were good (2014: $\chi^2(8)=279.1$, p<.001 & 2015: $\chi^2(10)=446.5$, p<.001)

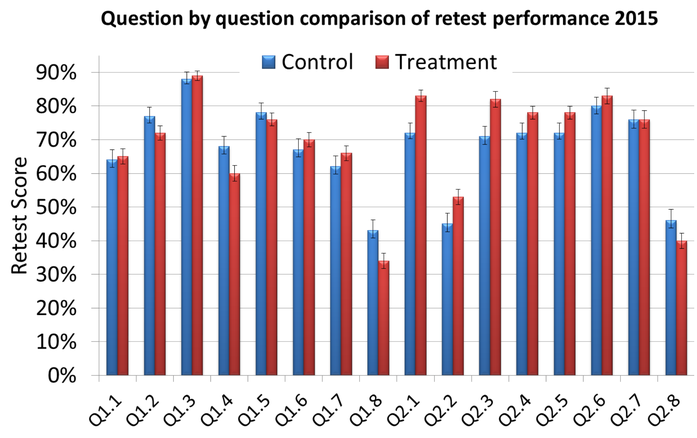
Longer timeframes (retest given with 4-7 weeks of *1st* midterm)

- For 2014, no statistically significant predictive power for retest questions Q1.1-Q1.x.
- For 2015, treatment predicted success for retest questions Q1.1-Q1.x (β_3 = -.193, SE = .077, p = .011)
- The fits between the model and data were good (2014: $\chi^2(8)=145.2$ p<.001 & 2015: $\chi^2(10)=440.0$, p<.001)

Question by question comparison of retest performance

Note: Question numbering schemes from 2014 and 2015 are not the same





Matched question pairs

The midterm exam questions were designed to form matched near-transfer pairs with questions on the locally developed end-of-term course diagnostic

Question validation:

Diagnostic question validation via:

- Expert feedback and student interviews
- Classical Test theory analysis ongoing

2014 Exam question validation via:

- Four course instructors
- Graduate student TA feedback

2014	Similarity Rating (SD)	Exam Questions		Diagnostic (Retest) Questions
		Fraction Correct	Disc. Index, D	Disc. Index, D
Q1.1	3.29 (1.11)	.453	.341	.270
Q1.2	4.00 (0.58)	.474	.324	.315
Q1.3	4.71 (0.76)	.636	.440	.465
Q1.4	4.57 (0.53)	.744	.403	.387
Q1.5	3.14 (1.07)	.610	.490	.345
Q1.6	4.28 (0.49)	.820	.335	.405
Q2.1	3.71 (1.11)	.841	.231	.275
Q2.2	3.86 (1.46)	.634	.370	.200
Q2.3	4.86 (0.38)	.837	.167	.385
Q2.4	4.86 (0.38)	.626	.305	.432
Q2.5	5.00 (0.00)	.691	.399	.464
Q2.6	4.86 (0.38)	.284	.402	.500

Similarity index: 7 content experts rated each question pair using a 5-point system:

- 5: target the same application of the same concept
- 3: target different applications of the same concept, and
- 1: target completely different concepts.

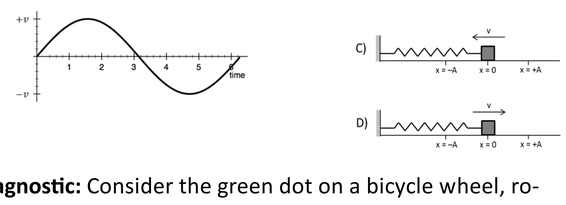
Discrimination index, *D*, measures how well the question discriminates between high-performing (top 21%) and low-performing (bottom 21%) students. An item having $D \geq 0.3$ is typically considered to have good discrimination (Day & Bonn, 2011):

- D = 1: All of the high-performing and none of the low-performing students answer correctly
- D = 0: High-performing and low performing students answer the question equally well

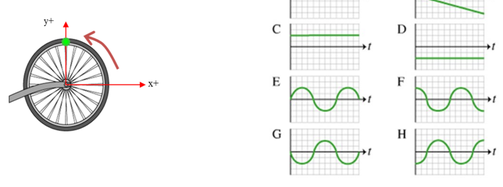
Example matched question pairs (2014):

Q1.5: Similarity rating = 3.14

Midterm: Given the plot of velocity as a function of time shown, which one of the following images best represents the situation at t = 0?

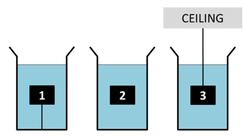


Diagnostic: Consider the green dot on a bicycle wheel, rotating at a constant rate in the direction shown. At t = 0 the green dot has the position indicated in the figure. Which of the graphs corresponds to the x-position of the green dot versus time?

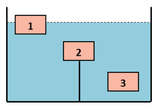


Q1.3: Similarity rating = 4.71

Midterm: Three identical beakers are each filled with the same amount of water and equal volume blocks placed in them. The figure shows the blocks at rest in their beakers. Block 1 is attached to the bottom of its beaker by a string and block 3 is hanging from the ceiling by a string. In each of the cases there is a non-zero string tension. Rank the buoyant forces experienced by the blocks, from largest to smallest.

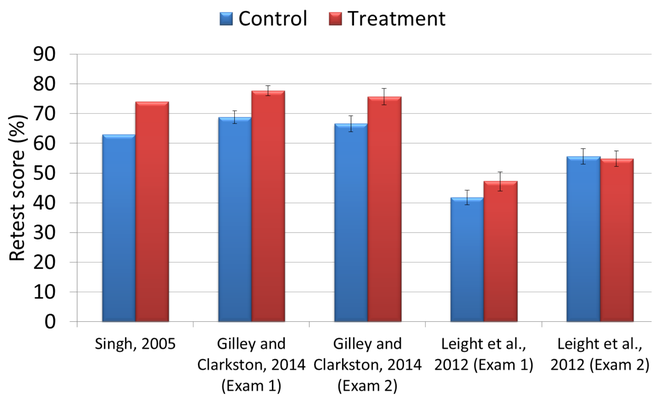


Diagnostic: Three objects having equal volumes are submerged in a fluid as shown. Object 2 is tethered to the bottom and object 1 is floating and only partially submerged. Rank the buoyant forces experienced by the blocks, from largest to smallest.



Results from previous studies

Similar studies in Physics (Singh, 2005) and Earth and Ocean Science (Gilley & Clarkston, 2014) showed improved learning from a collaborative group-exam treatment when the retest used *the same* questions as the initial individual test. A similar study in Biology (Leight et al., 2012) showed no improved learning on the retest.



References

- J. Day and D. Bonn, *Phys. Rev. ST Phys. Educ. Res.* **7**(1), 010114 (2011).
- B.H. Gilley & B. Clarkston, *J. Coll. Sci. Teach.* **43**(3), 83 (2014).
- H. Leight *et al.*, *CBE Life Sci. Educ.*, **11**(4), 392 (2012).
- C. Singh, *Am. J. Phys*. **73**(5), 446 (2005).