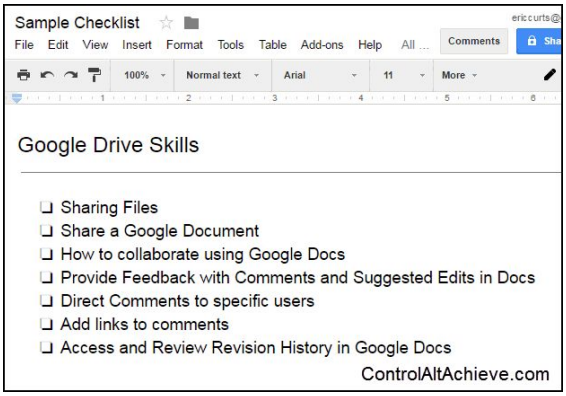
## Instructions for Using This Template

1. Go to File > Make a Copy. Do not edit the template.
2. Rename the File (yourlastname_digital_learning_workshop)
3. Type appropriate information below the pre-set headings. Information starting on the abstract page count toward the 2-page limit (single-spaced and font as given). Page 1 is only for organizational purposes.
4. When complete, go to File > Download as PDF to upload to EasyChair.




## Category

- ❏ Data-Driven Course Design
- ❏ MOOC Success Stories
- ❏ Blended/Hybrid Learning
- ❏ Applications of the Science of Learning (in Online and Blended/Hybrid Learning)
- ✓ Innovative Use of Digital Learning Environments (such as interesting uses of MOOCs, etc.)

## Title

Online quizzes to replace traditional paper assessment in introductory mechanics

### Author(s)

Byron Drury*, Ian Bouche*, Amelia Guttentag*, Sunbok Lee**,Michelle Tomasik*, Alexander Shvonski*, Chandralekha Singh***, David Pritchard*

### University or Organization

*Massachusetts Institute of Technology, **University of Houston, ***University of Pittsburgh

## Abstract

We administered frequent half hour algorithmically graded online quizzes in two introductory mechanics courses. We found that these online quizzes were as predictive of performance on a written final exam as traditional written quizzes and were more statistically self consistent, suggesting greater fairness. However, conceptual multiple choice questions on the quizzes behaved very differently between the two classes in terms of both self consistency and correlation with the final exam. We explore possible explanations for this effect.

## Introduction

Physics professors generally consider 'traditional' long, multi-step, open response paper examinations to be the gold standard for assessing students' physics skill. However, decades of physics education research have shown that such tests do not adequately measure students' conceptual understanding [1,2]. Further, hand grading of traditional exams consumes an inordinate amount of instructors' time and is inherently imprecise even when using a well designed rubric. How strictly and consistently a rubric is applied varies both between different graders and, over time, for individual graders. This grading error is unfair to students and detrimental to the purpose of accurately measuring students' (and instructors') performance.

With these and other issues in mind, we made a set of computer graded online quizzes for introductory physics and administered them to 120 students in two MIT classes ("8.011", n=15, and "8.01L", n=105). Our data suggest that online quizzes are more reliable than written quizzes and comparable in their ability to predict performance on final exams. There were notable differences between the two classes. While the multiple choice (MC) part of the quiz in 8.011 correlated far better with concept questions on the final than did other parts of the quiz, MC questions in 8.01L correlated poorly with all parts of the final and had low self consistency.

## Methods and Results

We generated half hour online quizzes using, where available, questions from research validated assessment instruments [3,4,5] and supplemented these questions with problems tested in large enrollment MOOCs. Questions were selected to span a typical week's topics and to have appropriate difficulty and discrimination (ability to distinguish high from low skill students). We used several question formats including multiple choice, numerical input, and symbolic (mathematical expression) input. They comprised a blend of "conceptual" problems requiring little or no calculation and analytical problems requiring mathematical manipulations. Written quizzes consisted of one, sometimes two, traditional problems.

The written and online quizzes (eleven in 8.011, seven in 8.01L) were given simultaneously; students had one hour to complete both parts and could divide their time between them however they saw fit. Each class had a three hour written final exam composed of concept questions and long problems akin to the written quizzes. Reliability of the quizzes was measured with Cronbach's α, a standard psychometric test for quantifying the extent to which a set of items (e.g. weekly quizzes) measures a single underlying trait. $\alpha=0$ indicates no correlation between items and $\alpha=1$ is perfect correlation; $\alpha>0.7$ is considered acceptable. We

evaluated correlations of the quizzes and their constituent parts (written [Wr], MC, online open response [OR]) with the final and its components (CONcept questions, TRADitional problems).

| | MC | OR | OL | Wr | Tot |
|---|---|---|---|---|---|
| α | 0.66 | 0.71 | 0.79 | 0.61 | 0.82 |
| CON | 0.81 | 0.49 | 0.82 | 0.37 | 0.83 |
| TRAD | 0.68 | 0.81 | 0.84 | 0.78 | 0.84 |

Table 2: Cronbach's α and correlations with the conceptual and traditional sections of the final exam for components of the quizzes in 8.011. OL is the online quiz (MC+OR), Tot is the entire quiz (OL+Wr).

| | MC | OR | OL | Wr | Tot |
|---|---|---|---|---|---|
| α | 0.48 | 0.78 | 0.81 | 0.60 | 0.83 |
| CON | 0.43 | 0.53 | 0.55 | 0.50 | 0.57 |
| TRAD | 0.38 | 0.52 | 0.53 | 0.59 | 0.61 |

Table 1: Cronbach's α and correlations with the conceptual (Con.) and traditional (Trad.) sections of the final exam for different components of the quizzes in 8.01L

Reliability of the online quizzes was better than that of the written quizzes in both courses, as shown in Tables 1 and 2. This supports the hypothesis that computer grading reduces the noise introduced by human graders. The correlation of online quizzes with the final exam did not differ significantly from that of the written quizzes, except for the conceptual portion of the final exam in 8.011, which we will discuss in detail. Correlations were lower in 8.01L than 8.011 for almost every pair of quiz and final components.

In 8.011 the written quizzes correlated poorly with the conceptual part of the final, while online quizzes correlated extremely well. Conversely, in 8.01L online quizzes had essentially the same correlation as the written quizzes with CON. This discrepancy is mainly due to the difference in the performance of the MC quizzes. Self consistency and correlations for the MC quizzes in 8.01L were strikingly lower than in 8.011.

While we cannot dismiss the possibility that the performance of MC questions in 8.011 was a fluke, the results comport well with previous research showing that concept questions measure aspects of knowledge which are not measured by typical calculational problems, implying stronger correlation of CON with MC than with OR or WR. In this light, the 8.01L results seem surprising. A vast body of research shows that active learning (as in 8.011) is far more effective at teaching concepts than lectures (as in 8.01L). Perhaps 8.01L students were guessing too much on the concept questions for them to be reliable measures of skill. However, the 8.01L students actually did quite well on the conceptual part of the final (compared to past MIT students given the same questions; we lack 8.011 data for comparison). Another factor is that most 8.011 students have already taken and failed an intro mechanics course at MIT, while 8.01L students are in their first semester at MIT and have little previous physics experience. Finally, students in 8.01L take the course "pass/no record", meaning that they do not get a letter grade on their transcript, just a "P" if they pass. This may result in some students who do very well on the quizzes exerting minimal effort on the final since they were by then confident in passing (or vice versa). However, it is not clear why this would disproportionately affect the multiple choice/conceptual part of the tests.

The data are unambiguous in supporting the use of online quizzes in place of paper assessment. Nevertheless, the details point to interesting avenues for further research. We have continued to administer both online and written quizzes and we are now analyzing data from further experiments to learn more about the relationship between different student knowledge structures, student backgrounds, teaching methods, and modes of assessment.

## References

[1] D. Hestenes and M. Wells. *The Physics Teacher*, 30(3):159–166, 1992.
[2] E. Mazur. *American Journal of Physics*, 67(4):359–360, 1999.
[3] A. Pawl, A. Barrantes, C. Cardamone, S. Rayyan, and D. E. Pritchard. In *Physics Education Research Conference* 2011, volume 1413 of *PER Conference*, pages 287–290, Omaha, Nebraska, August 3-4 2011.
[4] L. G. Rimoldini and C. Singh. *Phys. Rev. ST Phys. Educ. Res.*, 1:010102, Oct 2005.
[5] C. Singh and D. Rosengrant. *American Journal of Physics*, 71(6):607–617, 2003.