A mastery-based learning and assessment model applied to 3.091r (*Introduction to Solid-State Chemistry*)

Michael J. Cima

Summary

Software assessment tools developed in the context of MITx were used to implement a mastery-based learning and assessment system for 3.091r, an introductory chemistry course, during the fall of 2013. Students were required to pass a minimum number of assessments but were allowed to repeat each assessment as often as needed to pass within a 14 day period. A new problem was drawn from a database each time. Assessments were performed in a proctored environment. Lectures and recitations were carried out as in previous years of the course, other than an 80% class attendance was required (there had been no previous policy on attendance). No homework or written exams were assigned. The primary outcome measure was achievement by the 2013 class in comparison with the fall 2012 class. Outcomes were an average four-fold higher in 2013 over 2012 across eleven of the twelve outcomes and excludes the largest (a factor of 48-fold higher). The smallest increase was 30% and the largest was 4800%. The 2013 3.091r assessment system made a higher achievement requirement accessible to a broader number of students. Those students grappling with learning a new topic received multiple chances to demonstrate their competence. The lower achievement requirement and partial credit in 2012 may have been a disincentive for greater achievement. The fall 2013 3.091r class demonstrated levels of achievement most likely never before obtained by earlier 3.091 classes.

3.091 Background:

"Introduction to Solid-State Chemistry" (3.091r[1]) seeks to help students learn chemical principles through study of the solid-state. The underlying theme for the course is that the chemical bond determines properties. The outcome metrics look very similar to a traditional university chemistry course (chemical energy balance, bond strength, valence, acid-base equilibria, etc.). The principal differences are the examples used to reach those outcomes. 3.091r has satisfied the chemistry General Institute Requirement (GIR) at MIT for decades because the outcomes metrics are so similar to any first year chemistry course.

The course's emphasis is on linking basic concepts to applications as an engineering course would. Frequently taken during the freshman year, 3.091r may be the first engineering course that students take at MIT. The course differs from the traditional university chemistry course in more than just the reference to materials properties. Specifically, the focus is on what students *can do* rather than what they *know*. Assessment problems reflect that focus, as they are heavily weighted toward calculating a result rather than regurgitating an answer.

[1] This report refers to two versions of 3.091; 3.091r and 3.091x. 3.091r is the residence based version of *Introduction to Solid State Chemistry*. 3.091x is the on-line (MOOC) version of the course that is available on the edx platform. 3.091x has been offered to the worldwide audience three times since October 2012.

3.091r has not had a formal text for many years. The course used a combination of a standard university chemistry text, course notes written for the class, and a reader of individual chapters from several other texts. None of these sources covers the detail that is actually covered in the class, meaning that lectures either overlook material or have to go through much more detail than can really be assimilated in the standard 50-minute lecture. The lack of a text also has a negative impact on the number of examples that can be covered in lecture, the number of questions that can be responded to, the number of questions that can be put to the class, and the number of demonstrations that can be performed.

Assessment in 3.091r was traditionally composed of three monthly fifty-minute tests, a three-hour final, and approximately ten single-problem homework quizzes taken in recitation (ten minutes). The students were allowed a single sheet of notes for each cumulative exam (monthly tests and final). The value of each of these assessments on the overall course grade varied slightly over the years but the typical distribution was 30% final exam, 20% on each of the monthly tests, and 10% on the combined homework quiz scores.

Achievement was assessed on a partial credit basis for every problem that the student encountered. The idea was that student achievement could be assessed by *how close they got to a correct answer*. Passing the course required that the student achieve 50% of the total possible points. Freshmen at MIT take courses on a Pass/No record basis in the fall, so grades are not really an objective for most of the students taking 3.091r (the course is also offered in the spring but typically to a much smaller number of students). Obviously, these two policies mean that (freshman) students that have done well on the monthly exams can effectively opt not to take the final exam. Not taking the final was an infrequent, but not unheard of, occurrence.

The on-line version of 3.091 (3.091x) was first offered beginning October 15th 2012. The course was purposely started six weeks later than the residence-based 3.091r course so as to not expose the fall 2012 class to the on-line material. Outcomes for the residence-based class were going to be compared to outcomes for the on-line students. That outcome comparison suggested that the assessment model used in the residence-based class (3.091r) was deficient as a way to assess student achievement. Specifically, the hypothesis was made that the *cumulative fixed-time* exams either did not actually measure achievement or were not conducive to learning.

Changes for the Fall 2013 Class

The 3.091r fall 2013 class included several changes. First, the text for the course was replaced with the on-line material from 3.091x. Secondly, student assessment was changed to a mastery-based combined education and assessment model described below. The idea was to use tools built for the on-line class in a way to get frequent and rapid feedback to students about their progress. Lecture and recitation formats remained the same. The primary outcome measure was a comparison of the 2012 outcomes (as measured on the 2012 final) to the 2013 outcomes[2]. The plan was approved by MIT's Committee on

[2] The 2013 and 2012 class compositions were similar. There were 327 students registered for the 2012 class at the end of the term; 299 freshmen and 28 upperclassmen. The 2013 class had 361 students, of which 299 were freshmen and 62 upper classmen. A self-reported survey of the 2013 class taken at the beginning of the term was

the Undergraduate Program (CUP). Discussion with staff at MIT's Committee on the Use of Humans as Experimental Subjects (COUHES) indicated that the plan satisfied the administrative exemption criteria so no signed consents were required from the students.

The 2013 students took up to 37 proctored on-line assessments over the course of the semester. A passing grade for the course only required successful completion of 27 assessments. Flexibility was given to the student as to which assessment(s) to take in each unit of material with a minimum number of successful assessments required for each unit. The topics of the course were divided into 14 units. The individual assessments and units are described in the table below. There were no monthly exams and no cumulative final was given.

Each assessment consisted of a single problem and represents a learning outcome measurement. Assessments were administered in an Athena cluster in which a proctor was present. In addition, a TA was present in a neighboring room to be available for questions. The room was available from 7 pm to 10 pm Sunday through Thursday and three hours midday on Friday. Students were allowed as much time as they wanted during the open period each day, but were only allowed to take each available assessment once per testing period. There was a specified two-week time window to complete each assessment. If an assessment was not answered correctly, the student could take it again as many times as they wanted within the time window, but there was a calendar day lockout between assessment attempts. The student identified themselves to the proctor for each assessment and identification was verified using their MIT issued student ID card. The proctor would then set permissions so that the student could take the desired assessment. There was no web access and no notes used during the assessment. Each student did have access to a calculator, periodic table, table of physical constants, and a standard set of brief notes (the "3.091 Pocket Guide").

Each assessment problem was randomly selected from a group of many related problems in what is called the "3.091r Problem Bank". The problem bank was essentially a collection of independent databases concerning each assessment. Each database consisted of ten to thirty problems that were based on many years of 3.091r exams. The problem bank consists of both floating-point response and multiple-choice questions. Floating-point response assessments are 68% of the problem bank and require that the student calculate a numerical answer. For each assessment attempt, students are allowed three attempts to answer the question to give the student a chance to correct simple calculation errors. A similar approach is taken in many MOOC exams. The multiple-choice assessments consisted of multiple questions and only one response was allowed. Every attempt was made to make sure the difficulty level of each problem was similar to problems on written exams from the past. Parameters for many of the individual problems are randomly selected. Thus, each student sees a new problem each time they sit down to see an assessment. The lock-out period was desired because it was thought to prevent students from searching for the easiest problem they can find in the Problem Bank without a significant time or effort penalty.

used to gain some understanding of the students' backgrounds. They reported that they had an average of 1.63 years of high school chemistry (stdev 0.85 year). Eight students had no previous exposure to chemistry and three students had taken the 3.091x on-line course prior to coming to MIT.

Defined dates were specified for completion of each unit. This policy was an attempt to keep the class more or less on the same topic, yet give some flexibility to the student. The two-week time window for completion of each assessment was implemented by allowing an automatic one-week grace period beyond the due date[3]. Secondly, extensions were liberally granted. A total of 175 time extensions were allowed. Thus, about half the students took advantage of a time extension.

Week	Lecture Date	Lecture and assessment titles	Assignments	Due date
Week 1	Sep. 4, W	1: Why Solid-State Chemistry		
	Sep. 6, F	2: Modern Chemical Concepts and Periodicity of the Elements	Unit 1 (must pass 2)	Sep. 14
Week 2	Sep. 9, M	3: The Electron and Light		
	Sep. 11,W	4: Atomic Structure	Unit 2 (must pass 2)	Sep. 21
	Sep. 13, F	5: Wave-Particle Duality		
Week 3	Sep. 16, M	6: Spin and the Multielectron Atom		
	Sep. 18, W	7: Electron Transfer and Ionic Bonding	Unit 3 (must pass 2)	Sep. 28
Week 4	Sep. 23, M	8: Covalent Bonding		
	Sep. 25, W	9: Periodic Trends and Bonding		
	Sep. 27, F	10a: Molecular Orbitals	Unit 4 (must pass 2)	Oct. 5
Week 5	Sep. 30, M	10b: Hybridization and Molecular Geometry		
	Oct. 2, W	11: Intermolecular Forces and Materials Properties		
	Oct. 4, F	12: Reaction Kinetics	Unit 5 (must pass 2)	Oct. 12
Week 6	Oct. 7, M	13: Reaction Mechanisms		
	Oct. 9, W	14: Band Theory of Solids	Unit 6 (must pass 2)	Oct. 19
	Oct. 11, F	15: Band Gaps and Optical Properties		
Week 7	Oct. 16, W	16: Conductivity of Semiconductors		
	Oct. 18, F	17: Crystal Structures	Unit 7 (must pass 2)	Oct. 26
Week 8	Oct. 21, M	18: X-Rays and their Generation		
	Oct. 23, W	19: Diffraction and Braggs' Law		
	Oct. 25, F	20: Bonding and the Mechanical Behavior of Solids	Unit 8 (must pass 2)	Nov. 2
Week 9	Oct. 28, M	21: Defects in Solids		
	Oct. 30, W	22: Plastic Deformation and Defects		
	Nov. 1, F	23: Amorphous Inorganic Solids	Unit 9 (must pass 1)	Nov. 9
Week 10	Nov. 4, M	24: Properties of Glasses		
	Nov. 6, W	25: Diffusion in Solids	Unit 10 (must pass 2)	Nov. 16
	Nov. 8, F	26: Time-Dependent Diffusion		
Week 11	Nov. 13, W	27: Solutions and Chemical Equilibrium	Unit 11 (must pass 2)	Nov. 23
	Nov. 15, F	28: Equilibrium Between Phases		
Week 12	Nov. 18, M	29: Multicomponent Phase Diagrams		
	Nov. 20, W	30: Chemistry of Carbon	Unit 12 (must pass 2)	Nov. 30
	Nov. 22, F	31: Polymer Synthesis		
Week 13	Nov. 25, M	32: Polymer Properties		
	Nov. 27, W	33: Surface Energy and Surfactants	Unit 13 (must pass 1)	Nov. 23
Week 14	Dec. 2, M	34: Molecular Aggregation		
	Dec. 4, W	35: Acids and Bases	Unit 14 (must pass 3)	Dec. 11
	Dec. 6, F	36: Amino Acids and Protein Synthesis		
Week 15	Dec. 9, M	37: Protein Structure and Structural Carbohydrates		
	Dec. 11, W	38: Perspective: chemical Bonding and Materials		
		Total Assessments required to pass	27	

[3] The one exception was assignments coming due in the last week of the course. Institute rules dictate that all work has to be completed by the last day of class.

Comparison of outcomes between 2012 and 2013

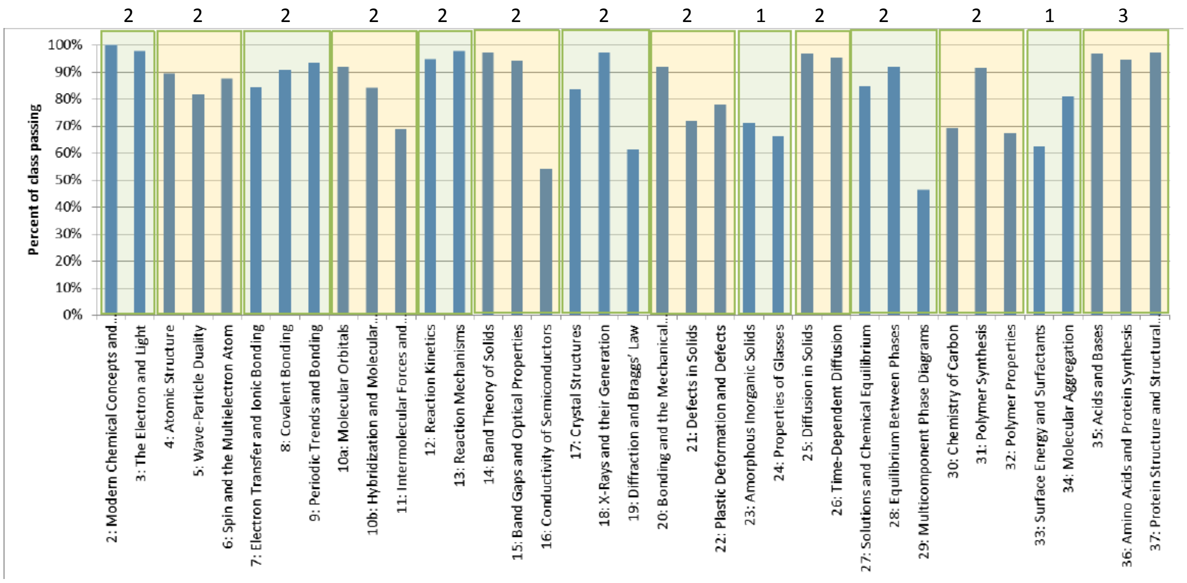
The 2012 final consisted of twelve problems. The individual problems were not cumulative and did not require synthesis among the different aspects of the class. This was consistent with past 3.091r finals that typically consisted of individual problems that might have existed on any of the monthly exams. Increased emphasis was placed on material that had not yet been examined on the monthly exams or material later in the course that had not had an examination opportunity. A one-to-one correspondence could be made for each of the twelve 2012 problems to individual learning assessments in 2013.

Similar to prior years, partial credit was given for each problem on the 2012 exam. The percent of the class getting a given problem completely correct was on average 8.2%. The percent of the class getting at least 80% of the credit for a given problem was on average 20%. The percent of the class getting *no problem completely correct* was 49%. This distribution of achievement is similar to my experience on past 3.091r exams. Partial credit averages on exams have typically been in the 60% to 70% range[4]. Only 50% of the total points are required to pass the course. The distribution of achievement in combination with the required points policy means that it was a very frequent occurrence to pass 3.091r *having never done a single problem completely correctly*.

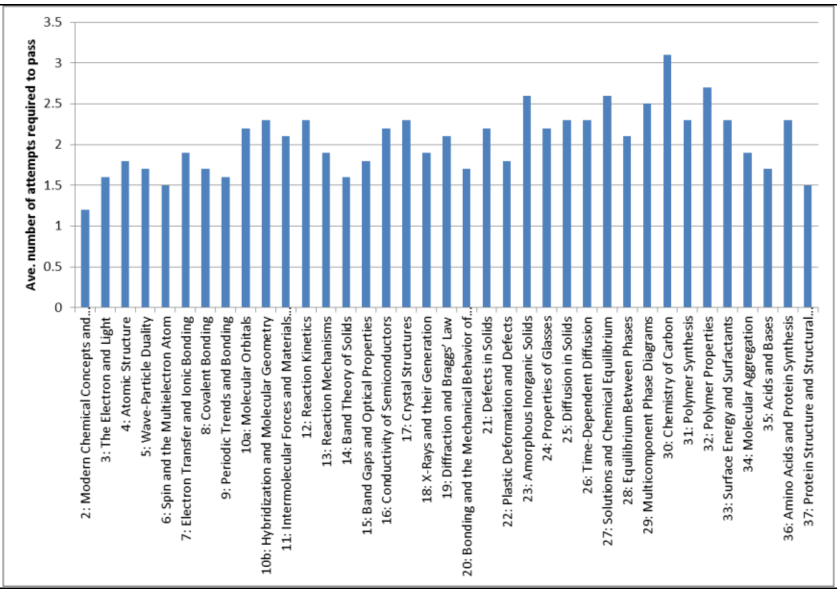
The record of achievement for the 2013 assessments is shown on the next page. Greater than 50% of the class has completed a problem in each assessment completely correctly. The average across all assessments is 84% achievement. The average number of assessments passed per student was 31 compared with the minimum required to pass of 27.

The average number of attempts required to pass each assessment is also shown on the next page. Two attempts were required on average for all the assessments, with a range of 1.2 to 3.1. The difficulty varied with each assessment in a manner that was similar to past experience with the course. The first third of the course is a review of AP chemistry that many students had taken in high school. The latter two thirds of the material is new to most students and this seems to be reflected in the increasing number of attempts required to pass.

[4] The 2012 final was slightly more difficult than past final exams. The average score was 109 points out of a total possible of 200 (55%) with a standard deviation of 26.7 points.



Record of achievement for 361 students registered at the end of the class in fall of 2013. The colored bars reflect the 14 individual units, each consisting of several assessments. The 361 students include a small number (3%) who had not yet formally dropped the class but had stopped taking assessments and attending class at various points during the term. Each unit is centered on a topic and consists of several assessments (two or three). The numbers above each bar indicate the required number of assessments in each unit. That is, students are frequently given a choice of the number of assessment required to pass a given unit (say two out of three). A low passing rate for a given assessment indicates there was a preference among students to take other assessments in that unit.



Average number of attempts required to pass each assessment. The overall average was 2.0 attempts per assessment.

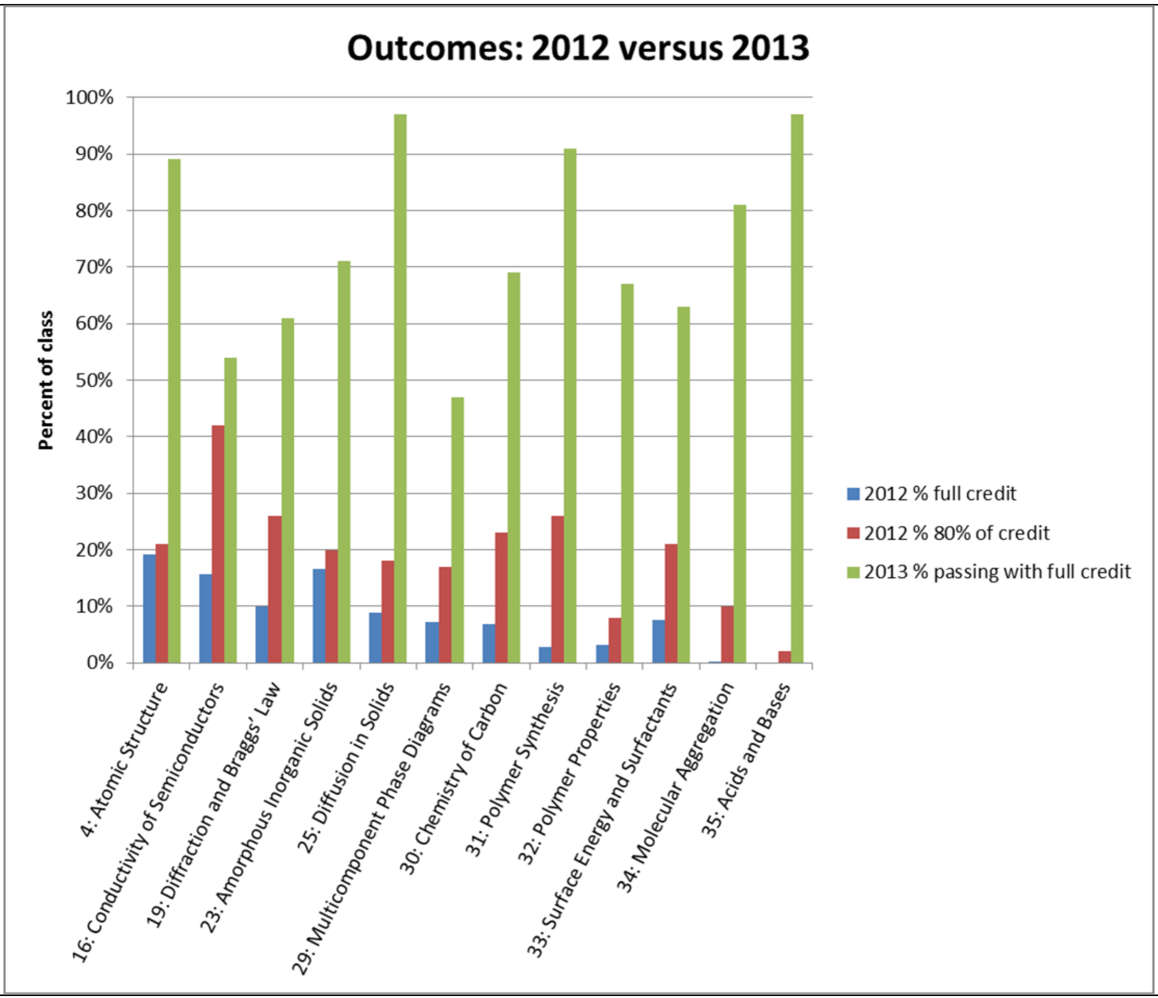
A comparison between the 2012 outcomes (as measured on the 2012 final) with the corresponding 2013 assessments is shown below. Each of the problems on the 2012 final has a corresponding assessment in 2013. Thus, the graph below compares level of achievement on each problem on the 2012 final with the level of achievement on the database of problems for the corresponding assessment.

Several points needed to be considered in the interpretation of this graph. It compares achievement on one problem in 2012 with achievement on a collection of problems in 2013. Each assessment database necessarily contains problems with a distribution of difficulty. Care was taken to try to avoid a wide distribution of difficulty and all the problems in the database were adapted from past 3.091r exam problems. It is, therefore, striking that the level of achievement is two to eight times higher in 2013 for ten of the assessments. The smallest increase was 30% larger achievement in 2013 for "conductivity of semiconductors". One of the outcomes (acids and bases) was forty eight times higher. Only 2% of the 2012 class was able to get at least 80% of the points on an acid-base problem. The 2012 students either did not know how to solve the problem, did not have enough time, or decided that they could pass the course with the points they had already accumulated. The situation was very different in 2013. 97% of the 2013 class was able to complete an acid-base problem completely correctly.

The additional difference is that the 2012 achievement was measured with partial credit while no partial credit was given in 2013. Minor calculation errors typically received 80% of the credit while conceptual errors received less credit. Student exams with problems completely blank were not uncommon. It is possible that a student was capable of doing the problem but ran out of time. Additionally, the policy of only requiring 50% of all points in the class to pass might have caused some students to opt out of finishing the exam. They really did not need to do more, once they felt they had reached that level of overall achievement. Thus, the 2012 measures probably bias the results to lower values.

The requirements to pass in 2013 probably pushed up the level of achievement. The 2013 student was given up to three attempts on a given floating point problem (to allow for possible minor calculation errors) and only one attempt for multiple choice problems. A minimum of 27 assessments had to be passed out of a possible 37. The choice was constrained, however, among the fourteen units. Each unit had to be passed. Thus, the bar for the 2013 students was higher since the number of assessments required was 73% (27 out of 37) and the degree of choice was lower. The 2013 students had only choice of assessments within each unit. There is evidence, discussed below, that as much as 40% of the class was optimizing their effort just to pass the class. The apparent increase in requirement most likely affected this group, but probably only at the level of change in baseline requirements, 73%/50% or a 46% percent improvement.

It is difficult to attribute the increase in 2013 achievement to any single cause. Undoubtedly, the known improvements in achievement found in earlier mastery-based education models played a role. More time to attempt a given assessment and the choice of when to take an assessment surely helped students. A change in requirements in 2013 over 2012 could have contributed. It is difficult to imagine, however, that the change in 2013 requirements can account for four fold increases in outcomes over 2012. The 2013 students were able to take any assessment multiple times through the online system with immediate feedback. The standard for 3.091r in previous years was to grade the monthly paper exams on the exam day but feedback was either one or two days later. A one or two day turn around was considered at least as fast as or better than other GIR courses at MIT. The frequent testing and frequent feedback opportunity experienced by the 2013 class is like nothing we have used before and is probably a major factor in these results.[5]



Direct comparison of the 2012 outcomes with corresponding outcomes for 2013.

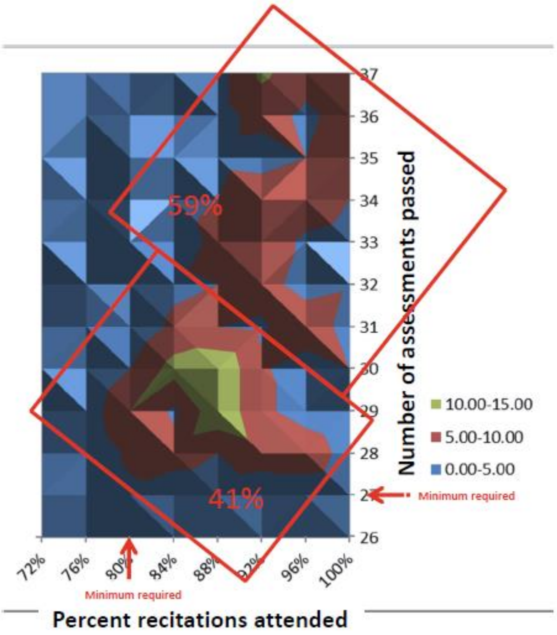
Students in 2013 were required to attend 80% of lectures and 80% of all recitations. This attendance requirement was new to the course. It was primarily motivated by an abnormally large upperclassmen registration at the beginning of the course. The initial registration for large classes at MIT changes on an hourly basis in the first week of class but 394 was an approximate value on the first day of class. Nearly one hundred upperclassmen were among those registered, which was significantly higher than previous years. Dialog with these students indicated that many of them actually planned to only take the

[5] See for example, “Classrooms –goals, structures and student motivation”, C. Ames J. Educational Psychology, Vol. 84, Issue: 3, Pages: 261-271 (1992). “Cognitive tutors: Lessons learned”, J.R. Anderson et al., Journal of the Learning Sciences, Vol. 4 Issue: 2, Pages: 167-207 (1995). “Focus on formative feedback”, V.J. Shute, Review of Educational Research, Vol. 78, Issue: 1, Pages: 153-189, (2008)

assessments and not attend any of the classes. The most common reason was because they were registered for other courses with conflicting schedules. MIT rules allow such conflicts in registration. Precise information on the degree of this issue was not available through the registrar's office, but it potentially affected as many as 20% of the students initially enrolled in 3.091r. Two problems were presented by this situation. The first was that it was represented to the CUP that the only changes being made were use of on-line content and the new assessment model. Specific discussion occurred about not giving GIR credit for completely on-line courses. Secondly, to continue including a group that was as large as 20% of the class that didn't have any contact time might complicate the outcomes comparison. A more thoughtful approach was undoubtedly possible, but the decision had to be made in the first week of class as required by MIT rules. After announcement of the new attendance requirements, the registration soon settled to 361, which included 62 upperclassmen.

Attendance was straightforward to implement in recitation, but the lecture requirement posed some challenges. It took several weeks to establish a card reading system so that student ID cards could be read and recorded as students entered the lecture hall. Records could only be obtained on the last 32 of the 38 lectures. Lecture attendance records probably overestimate attendance. Anyone requesting an excused absence ahead of time was given one. Some students noticeably scanned their ID and immediately left the room. Some students sat behind AV booth during lecture and did other work. These problems could be rectified should lecture attendance be used in assessment in the future. Attendance remarkably only affected the grade of a very small number of students. Less than 3% of the class failed due to not meeting the attendance requirement. Many of these students had probably stopped attending the class but had not formally withdrawn.

Shown below is a plot of recitation attendance versus total number of assessments passed. The minimum attendance of 80% is indicated as is the minimum number of assessments (27). Two features are worth noting. There appears to be an association between recitation attendance and number of assessments passed. Students that had attended more recitations also tended to pass more assessments. There is also an obviously higher concentration of students around the minimums of required attendance and successfully completed assessments. These students were exhibiting an optimization behavior by trying to just meet the requirements to pass the course. About 41% of the students were in the apparent just-pass category while 59% of the students exhibited much greater achievement with an average of 31 assessments completed successfully.



The number of students versus their recitation attendance and number of assessments successfully completed.

The course passing rate remained unchanged from previous years. The grade equivalent for passing is receiving a C or greater (for the pass/ no-record system at MIT, P/NR). The percent of the class getting a D or lower is the following for each of the indicated years so the typical passing rate is around 95%.

2009	2010	2011	2012	2013
8%	3%	4%	6%	5%

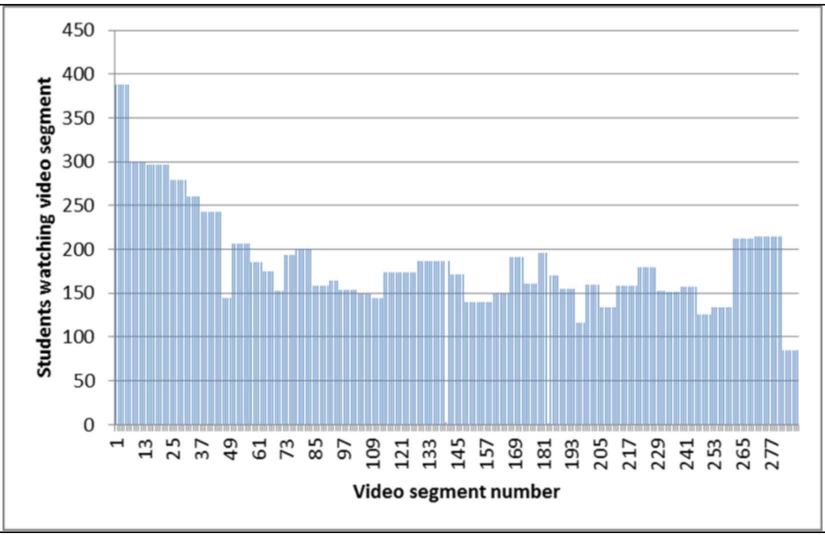
Use of on-line content and student behavior

The on-line content for 3.091r consists of 38 learning sequences (one for each lecture). Each learning sequence is composed of short video elements (either lecture segments or screencasts) and practice problems. Each week of learning sequences includes an additional number of practice problems. There were a total of 288 video segments and 337 practice problems made available over the semester. About half of the students made use of the on-line content during the fall of 2013. Shown below is the number of unique users for the video segments within each of the 38 learning sequences. A second figure below is the number of unique users for the practice problems. Usage of both starts high but then drops to approximately half the class. It appears that usage of practice problems persists somewhat longer than usage of video, but both end up at roughly the same level.

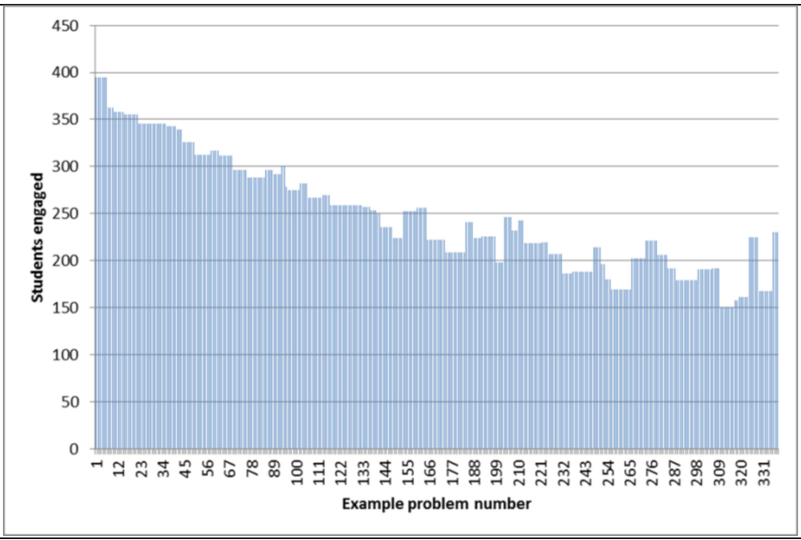
It is difficult to know whether 50% usage is high or low as there is nothing to compare with. The effort and cost to produce the on-line content is significant. Initial disappointment at the usage changed when several undergraduates remarked how surprised they were that it was so high. Their comparator was usage of conventional textbooks. An initial literature search turns up little to give guidance on actual usage of physical textbooks in a university setting.

If students are not preparing for assessments by using practice problems, what are they doing? The answer appears to be that they practice by doing assessments. While the average number of attempts required to pass is 2.0, the variation is quite wide. The standard deviation is approximately 1 so frequently students were making three or more attempts to pass. Another common behavior was that most students attempted all the assessments they could get access to in a single session. For example, if a unit consisted of three assessments, some students would try all the assessments until they could pass the total number required. If only two were required, they would stop at that point. It appears that quite a number of students were optimizing *by studying in the assessment room*. If one must practice problems, one might as well do so under conditions where one might also pass the assessment. There was no penalty for repeated attempts except the lockout period of one calendar day. In addition, there was always a teaching assistant in the next room to answer questions.

The net result was that the assessment room took on the ambiance of a study hall or study lab. The excess usage of the facility posed some resource challenges but may have contributed to the high achievement of the 2013 class. I have worked with 3.091r on and off over my twenty eight years at MIT. My impression has always been that student success with the course was directly related to practice solving problems. Creating an environment where practicing problems is encouraged appeals to my common sense but is difficult to quantify.



The number of unique students watching video segments within each learning sequence. There are 288 video segments in entire course divided among the 38 learning sequences. The video segment number grows in approximately chronological order. Usage drops to between 150 and 200 students over the course of the term.



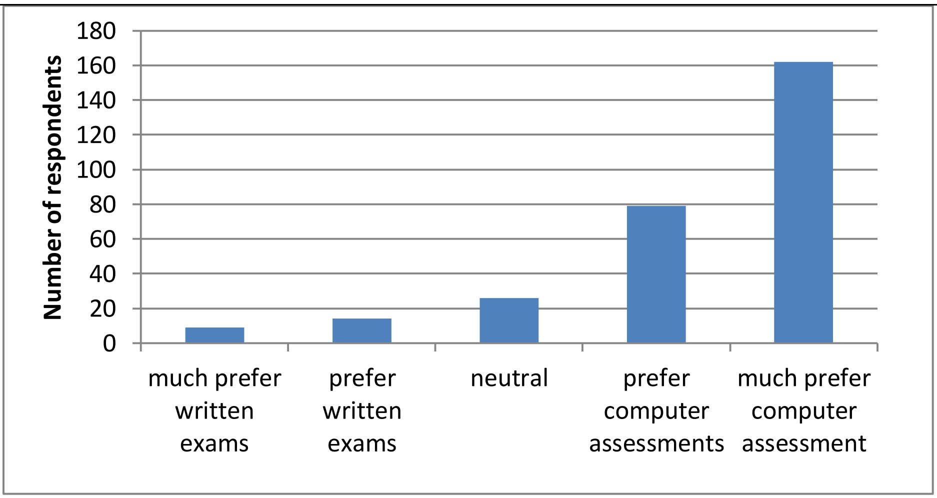
The number of unique students using on-line practice problems. There are 337 example problems within the on-line material. The example problem number grows in approximately chronological order. Usage drops to between 150 and 200 students over the course of the term.

Student perspective

Student comments on the course were collected via three different surveys. These surveys were taken under conditions where the identity of the respondent could be protected. The first survey was conducted during the sixth week of class and performed on the Survey-Monkey™ platform. The class was stratified by freshmen and upperclassmen and identical surveys were administered to each. Another Survey-Monkey™ survey was administered in the last week of class and similarly stratified among freshman and upperclassmen. Finally, a web-based standard MIT course evaluation was administered to all students enrolled during the last week of class. The latter two surveys were designed to complement one another.

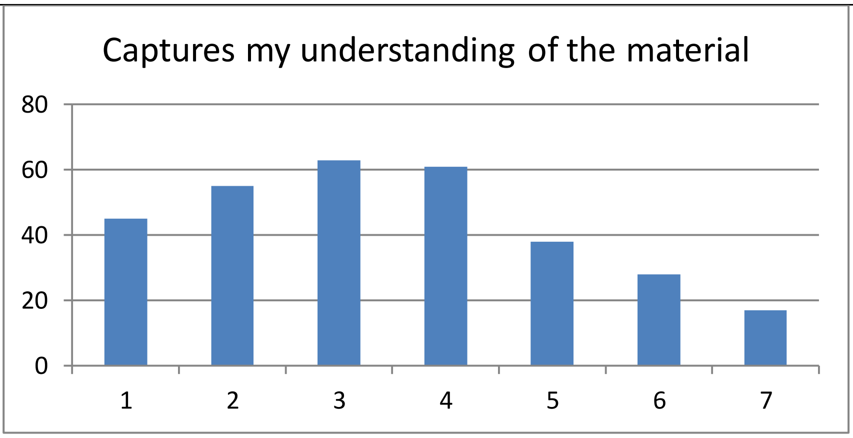
The timing of the first survey was selected to be given at a time when many of the students were taking, or had just completed, standard written exams in the other classes. The 3.091r students had just completed the unit 3 assessments. Thus, the timing made a question about their comparative experience appropriate. Shown below is their response. The response was overwhelmingly in favor of the computer assessment model. This result was somewhat surprising as the assessment system experienced a number of bugs in the first few weeks of the term. These included corruption of an entire assessment database by seemingly minor coding errors in one problem. The system never automatically transferred assessment completion data to Stellar™ (MIT's course management software). An improvised manual method had to be implemented where an administrative assistant would manually enter completion data from the night before. Students could not confirm their completion of an assessment until the next day which created some confusion at times. We had not fully implemented a problem debugging and validation procedure at the beginning of the term. It took at least four weeks to get team and process in place to validate the solutions to all problems. The students were given frequent updates on the staff recognition of issues and what was being done to correct the situation.

This may have offset some negative feelings toward the assessment model. Most problems seemed to have been addressed by mid-semester and a routine set in.

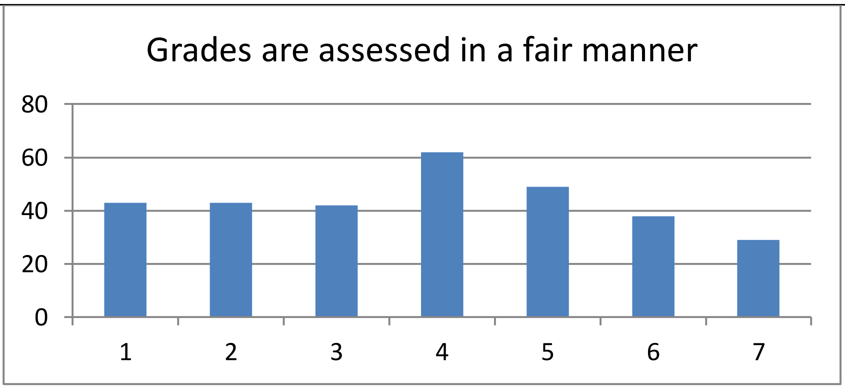


Results of sixth week survey in response to the question "*You're 21% completed with 3.091 with the completion of unit 3 this week. Normally, 3.091 would have had a 50 minute exam this past Monday or Wednesday worth 16% of your grade. Rank how you feel about the computer assessments relative to a written 50 minute exam at this point of the course. (Select one)*" 290 respondents.

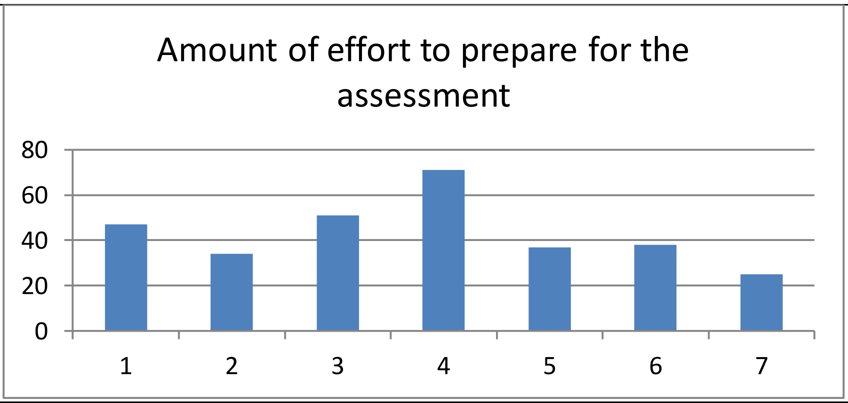
The end of term survey tried to explore whether students felt that the 3.091r assessment model was differentiated in any way from the conventional testing performed in MIT GIRs. Students were asked to compare whether it captures their understanding of the material. Secondly, they were asked whether they thought grades were assessed in a fair manner compared to conventional tests. Lastly, they were asked whether the amount of effort to prepare was similar. Students were asked to score their response on a 0 to 7 scale (1 compares poorly to other classes, 4 about the same, 7 much better than other classes). The results were less easy to interpret than their preference described above. Shown below are the distributions from the three questions. All the responses produced an average slightly less than 4. Only the response on whether the assessment system captures their understanding produced a roughly normal distribution. None of these factors seem to explain the preference for the 3.091r assessment system described above.



Average 3.47
307 responding



Average 3.85
306 responding



Average 3.76
303 responding

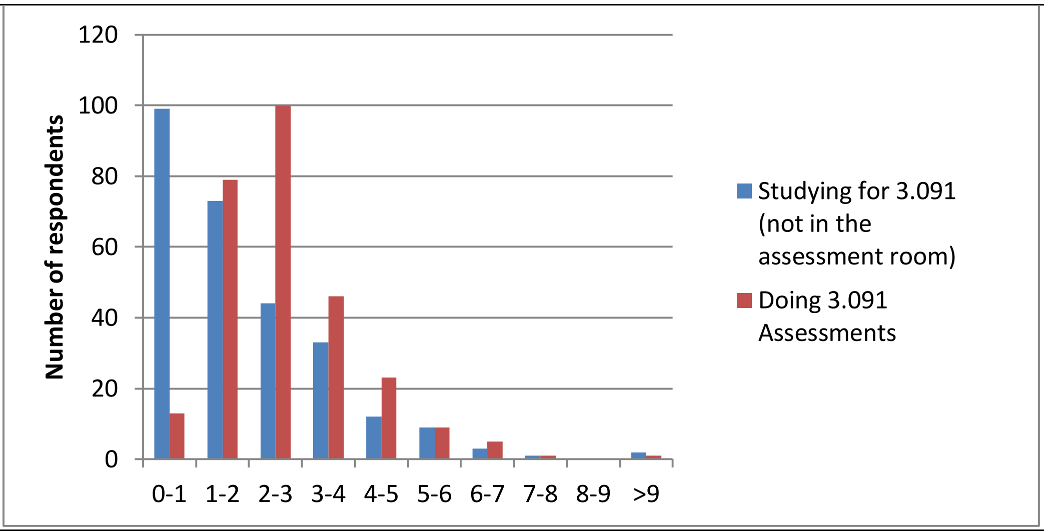
Total class response to the questions: Compare the 3.091 assessment system to the traditional exams that you have taken in your other GIRs. (1 compares poorly to other classes, 4 about the same, 7 much better than other classes), a) Captures my understanding of the material, b) Grades are assessed in a fair manner, c) Amount of effort to prepare for the assessment

3.091r is described as a 5-0-7 class at MIT. This designation means that it includes 5 contact hours (three one-hour lectures and two one hour recitations) and 0 hours in the laboratory[6]. Students should expect to spend 7 hours outside of class studying. Specific survey questions were asked at the end of the term on how much time students spent on the course. The Survey Monkey™ responses are shown in the figure below. The self-reported median time for studying outside the assessment room was 1.5 hours. The median time in the assessment room was 2.5 hours. If students did attend class they would

[6] 3.091r does not have a laboratory component.

have experience a time commitment of 9 hours per week. The MIT course evaluation also asks about time spent on the class but the standard questions are a bit confusing when applied to the 2013 versions of 3.091r. Nonetheless, the total time indicated had an average of 8.0 hours per week. The 2012 responses indicated an average of 8.3 hours per week. Thus, the new assessment model seems to not have added a time burden to the students.

The distribution of time outside of class is consistent with students studying in the assessment room. Freshmen reported spending more time in the assessment room than they did studying (2.5 hours compared with 1.5 hours). Half the students did not use the online material (videos and practice problems). Thus, their "studying" was really the time they spent in the assessment room. Assessment room records indicate about 1.3 hours per week per student. This time does not count time required to get to the assessment room and wait in line to check-in which could account for the difference with the self-reported time. The students passed an average of 31 assessments or about 2 per week. Thus, the time to pass an assessment was 0.65 hours on average.

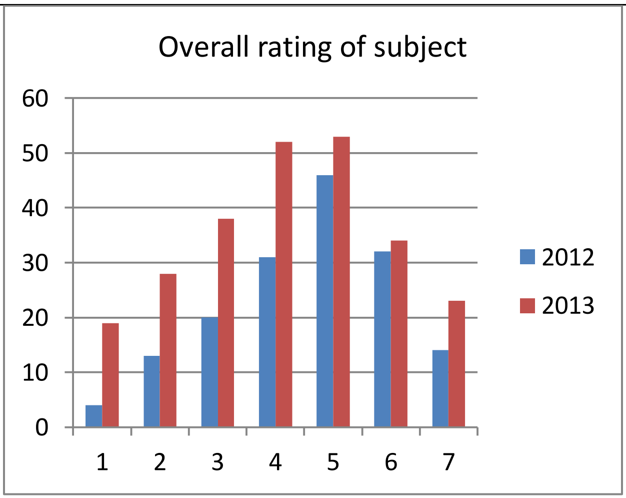


Self-reported freshman-only response to the question "*How many hours did you spend per week on the following? a) studying for 3.091 (not in the assessment room), b) doing 3.091 assessments*" The median self-reported time for studying outside the assessment room was approximately 1.5 hours. The self-reported time in the assessment room had a median time of 2.5 hours[7]. 3.091 is a 5-0-7 class which means five hours of contact time (lecture and recitation), zero hours of laboratory, and seven hours of outside study. Thus, the median time spent on the course per week was 9.0 hours (5 + 1.5 +2.5). This is similar to what was reported in the 2013 MIT standard course evaluation which reported a total of eight hours per week on average total for the course. The difference seemed to be that students reported slightly less than five hours of contact time per week. The 2012 course evaluation reported that students on average spent 8.3 hours per week.

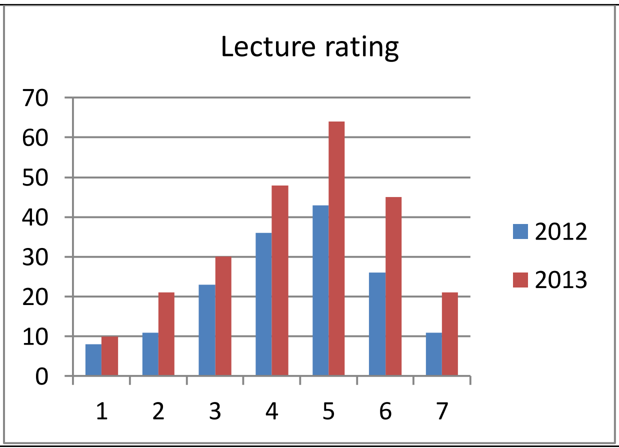
The MIT course evaluation asks for overall ratings of the subject and lecture. Ratings are on a 1 to 7 scale with 1 being the worst and 7 the best. The results are shown below. Comparison data for the

[7] Records for three weeks (starting 10/6, 11/3, and 11/17) show that total effort in the assessment room was an average 470 person-hours per week. There are 361 students which yields 1.3 hours per week for each student.

2012 class is also included. The average overall subject rating dropped somewhat in 2013 (4.2 from 4.6 in 2012) and the lecture rating increased in 2013 (4.5 from 4.4 in 2012).

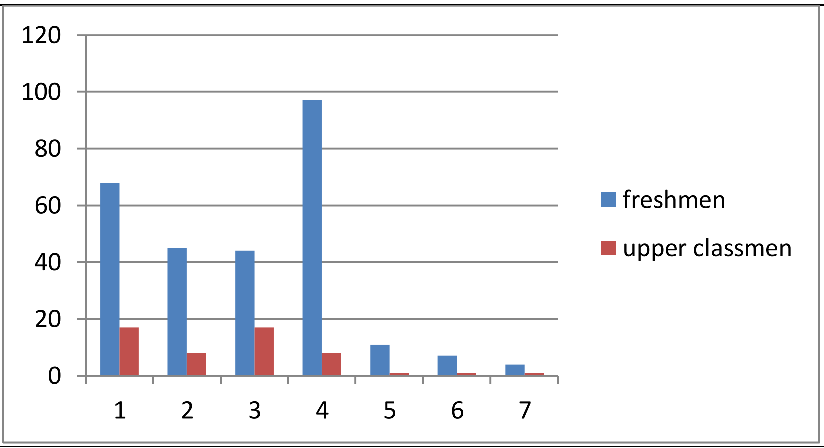


Overall subject rating (0 to 7 scale) for the two years 2012 and 2013. The average for 2012 is 4.6 with a standard deviation of 1.5. The average is 4.2 for 2013 with standard deviation of 1.7.



Overall lecture rating (0 to 7 scale) for the two years 2012 and 2013. The average for 2012 is 4.4 with a standard deviation of 1.4. The average is 4.5 for 2013 with standard deviation of 1.6.

Finally, students were asked whether the attendance requirement impacted their experience with the course. Shown below are the responses. Students generally felt that the attendance requirement negatively impacted their experience with a slightly more negative experience for the upperclassmen. Both distributions are not normal and exhibit bimodal behavior. The largest single group is freshmen that feel the attendance requirement had no impact on their experience.



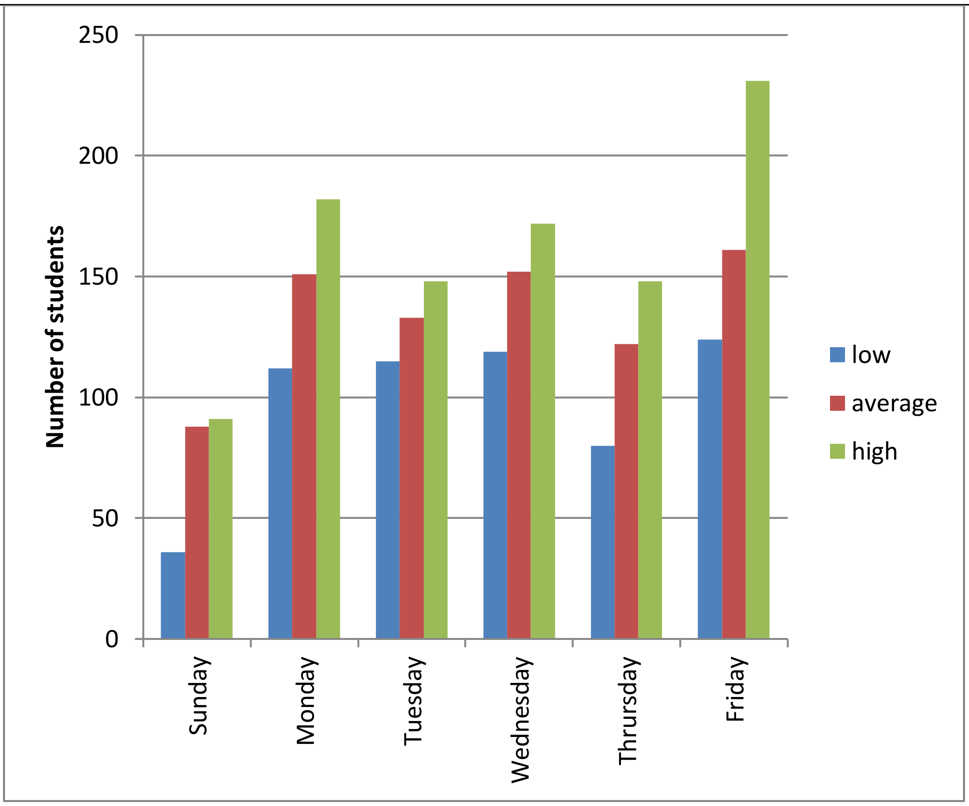
Response to the question "*Did the attendance requirement negatively or positively impact your experience with the course? (1 very negative impact, 4 did not impact, 7 very positive impact)*"

Resources

The most difficult resource decision concerned the assessment room. Space is limited at MIT and space dedicated to a single purpose is even more difficult to find. Our search was simplified by tring to reserve time after normal teaching hours and to use equipment that would permit the room to be reconfigured for its normal day use. The space was thus supplied with forty Chromebooks™ that could be locked away when the room was closed.

The afterhours approach introduced an additional complexity concerning institute rules about appropriate instructional hours. We asked for and received permission for the hours we settled on. There was some variation during the term but the basic timing was Sunday through Thursday 7 – 10pm and Friday from 12 – 3pm. We did successfully add hours near the end of the term as students were pressed to complete assessments on time.

Usage toward the beginning of the semester seemed to rise steadily until immediately before a due date (most due dates were by the end of the Friday session). Lines would form as students waited to verify identification and to wait for computer access. Students were encouraged to try earlier in the week to avoid lines and they seemed to respond. Shown below is usage data from three weeks beginning on Sundays, 10/6/13, 11/3/13, and 11/17/13. The average time for a given student spent in the assessment room was 35 minutes (time between checking in and checking out). Sundays usually experienced the smallest demand with peak usage less than 100. Peak usage on the other days of the week ranged from 150 to 230 students over each three hour period. The maximum possible usage was certainly reached at 230 students (230student/3hr x (0.5 hr/student) = 38 which is very close to the number of computers in the room). Not shown in the plot below is that usage was never uniform over the three hours. Peak usage was usually between 7 and 9 pm (Sunday through Thursday) and 12 to 1pm on Friday. Thus, lines could be observed during these times.



Summary of the number of students using the assessment room for three weeks beginning on Sundays, 10/6/13, 11/3/13, and 11/17/13. Hours of operation were Sunday through Thursday 7 – 10pm and Friday 12 – 3pm. The average time each student stayed in the assessment room was 35 min. The average number of students per session over all days was 134 with and stdev of 45. The peak usage for these three weeks was 231 students.

The additional resource required for this class was people. Eleven teaching assistants were required to teach recitations and be available so that at least one TA was present during the assessment times. Two proctors were presents during the assessment time in order to efficiently check-in students. There was a specialized team of two TAs that was dedicated to debugging problems with the Problem Bank. Finally, there was an entire team within the Office of Digital Learning that helped us with software infrastructure.

Recommendation

It is clear from the outcomes that the class of 2013 had far better achievement than the class of 2012. Despite the better outcomes, there is no evidence that students devoted more time to the class than in previous years. In addition, the requirement that students only pass when a problem is answered completely correctly added no undue time burden. These factors support continued use of the mastery model for 3.091r.

Student behavior indicated that they used the assessment room as much for studying as for completing tests. Class policies and organization going forward should somehow build on that model. We might, for example, revisit the 5-0-7 designation for 3.091r. A more appropriate division might be 5-3-4, if the time in the assessment room were considered a laboratory experience. The disadvantage of a formal "lab" is that it would be at assigned times and may pose a difficulty in scheduling with all the other freshman requirements.

Several important recommendations for further implementation of 3.091r are listed below.

- Software was not completely ready at the beginning of the term. Particularly frustrating to students and staff was the lack of an automatic link between the assessment systems and the Stellar™ course management system.

- The procedure for debugging problems needs to be formally implemented well before the class starts.

- The number of assessment problems should be increased in the Problem Bank. The students made two attempts on average to pass. Random selection of problems mostly worked but for those students having trouble and using more attempts, encountering the same problem did happen. Two features could change this. First is to generate more problems. The minimum number in each data-base was ten problems with some having more than thirty. We think the optimum might be twenty. In addition, rather than random selection of problems in the data base, problems should be assigned in a way that prevents repeat problems. That is, the system would track which problems have been assigned to each student.

- The proctor facilities need to be expanded in both number of terminals and available times. There is possibly an inverse relationship between total hours available and the increase in number of terminals required. If the available times were to remain the same, then we would need approximately 60 seats to handle peak load periods. Increasing the amount of time the facility is available could distribute the usage over more hours which would decrease peak demand.

- Check-in and check-out procedures need to be streamlined as both proctors and students report a bottle neck. A system should be devised that automatically checks in a student (i.e. allows access to the online assessment) with a swipe of their MIT ID card, then checks the student out upon a final swipe while leaving the classroom.

- Attendance should not be mandatory. Students should be incentivized to attend class by using their attendance record to obtain extensions on assignments.

- The MIT end of term regulations need to be modified to accommodate this type of evaluation model. Any evaluation of the material in the last week and a half of the course is subject to the rule that all assignments must be complete by the last day of class. This puts a great deal of pressure on students and the resources to accommodate their evaluation in just a few days. We had to change the calendar day lockout rule for the last week but we were still faced with a huge demand and unnecessary stress on the students.

Much discussion has occurred among students and faculty about the absence of a more comprehensive assessment as implemented in the fall of 2013. Some students and faculty feel that the act of studying for a final imparts learning that is not obtained when one studies for individual tests. I could find no data to support this position. What is certain is that 3.091r final exams have never been comprehensive. The finals have been composed of problems that are the same as those that appear on the midterm exams. Implementation of one or several comprehensive assessments would be relatively easy to do. This would, however, add some complexity for the student with an uncertain value.

Acknowledgments

This course would not have been possible without the dedicated work of quite a few people. I am very indebted to Kerri Mills for work well beyond her job description. Will Dickson and Mary Breton were invaluable for their work on the 3.091r Problem Bank and were always willing to offer fantastic advice. Ike Chuang and the ODL team made the entire infrastructure work and were a pleasure to work with. Jim Cain made the proctor room work. The TAs put in way more hours than they planned and I am very grateful for that.

Fall 2013 3.091r Team	
Course administrator Kerri Mills Course secretary Barb Layne Exterminators Mary A Breton William F Dickson Faculty and senior staff Niels Holten-Andersen David Paul TAs: Victoria Enjamio Zhaohong Han Donghun Kim Emily McDonald Kunal Mukherjee Maxwell Plaut Max Powers Christoph Sachs Wenhao Sun AV: David James Broderick Craig Milanesi	ODL: (Office of Digital Learning) Isaac Chuang Joe Martis Peter Pinch Caroline Soares Dan Carchidi Shelly Upton Benjamin Weeks OEIT (Office of Educational Innovation and Technology): James R Cain Proctors: Jenny Selvidge Jose Burgos Cecillio Aponte Amanda Evans Peter Augusciak Carolyn Joseph Selda Buyukozturk Emma Gargus Katelyn Rossick Elizabeth Murphy

Several comments were received from students (and one undergraduate proctor) in response to the report. These are included below.

Isaac Silberberg, 3.091 class Fall 2013
Hi Professor Cima,

Thanks so much for sending this out. I think another frustration for some people (myself included) is that the grading was rather vague about how many assessments it took to receive each grade. In this case, it did feel weird that a very high completion rate could still receive a B grade.

I am concerned that your report missed a significant benefit of the fall format though: stress and life balance. I am well-known to panic and be very nervous around exams, especially ones where I am unsure of what I will be asked to demonstrate. Many other GIRs, like 8.02, have tons of practice materials which helps compensate a little bit. No matter though, I still wind up with dreaded ³hell weeks² with 3 or 4 exams, sometimes on the same day. In these situations it¹s unlikely for me to perform well or actually learn all the material well. 3.091, on the other hand, was unlike anything else because I could simply focus on the material and "consumption smooth" (Econ major here) my workload to work around my other activities. If I seized up, or had a panic attack, I could come back and do the assessment again, which I often did. I could also plan to take a bit of a step back during weeks when I had many more exams in other classes, and refocus on 3.091 when I had more time. I implore you to fight for the continuation of the EdX assessment model because even if the data doesn¹t show a significant difference in outcomes, I think the reduction in stress is something that is incredibly valuable for a particularly marginalized group of students here at MIT.

Finally, I think that you could push the 3.091r experiment a bit further.
If there was no deadline on assessments, I think that this is what the future could really look like (especially as more courses adopt this model). Students could focus on one class and just zone in on it if they got in a groove, or set up a lighter week as their other commitments see fit.

In short, I was a huge fan of the 3.091r "experiment" in the fall. While far from perfect, I do think that it is the future of our education, and someday we¹ll see the near-elimination of semesters (and perhaps of grades) as the year begins to resemble IAP more and more, and classes that are offline are heavily discussion based or highly interactive, such as the 2.009 product design class. I don¹t know if any other professors or departments have approached you about expanding this to other courses (I know I¹m curious), but I think this is a hugely positive development in MIT¹s history and would love to be kept in the loop about it.

Thanks professor Cima,
Isaac Silberberg

Ariel Jacobs, 3.091 class Fall 2013

Thanks so much for the great semester!

Not mentioned in the report was the real reason, in my opinion, that Friday was the busiest day in the assessment room: It was the first day one could do all the assessments at once. 3.091 should consider creating a schedule that allows students to "pick a date as their week end". Instead of giving a deadline of Friday, if the deadline for an assessment was ~ 2 weeks from the lecture, students would be more likely to distribute among themselves the days on which they come in to do assessments, and cause a lesser load on Friday. I also think that a room closer to where most students spend their time might encourage others to visit on less convenient days.

I appreciate the recognition and appreciation for those who "studied in the assessment room". It helped me learn those things I was obviously supposed to understand from high school, and gain a good understanding of what I needed to know for the assignments,

Ariel Jacobs

Carolyn Joseph, Undergraduate proctor for 3.091 Fall 2013

Hi Professor Cima,

I've read through the report and I'm glad you sent it to me because now I can better see some of the advantages of this system. It is pretty amazing that students are getting 100% of the problem correct when they pass each assessment, and I see how that is a major improvement from the previous system. Another thing I really like about the online assessments is that it removes the time component from the examinations. I remember feeling very rushed in the 50 minute exams and thinking that if I could just take a little more time, I could have completed much more of the problems. I think there was a full problem I didn't even get to on the first exam. The comment about immediate feedback is also something I hadn't thought about.

I like your recommendations at the end, especially about improving the logistics of the assessment room. My main problems as a proctor were the long lines, inefficiency checking for IDs and manually granting access, and dealing with bugs in the system. Being able to automate the system with ID card readers would be great. Another problem was the ambiguity of the assessment room rules. Many proctors had different routines and policies (for example the extent to which they would help clarify a question or under what circumstances they would allow students to leave the room to see a TA.) I think it was confusing for students to have such different routines on different nights. I suspect that the low course evaluations have a lot to do with inefficiencies and frustrations with the assessment room.

I suppose my main concern with this new system is still the lack of any comprehensive assessments. Students are never held accountable for the information they learn aside from the single question they must pass about each topic. Even if students study for exams by cramming at the last minute, the exams still serve as incentive to understand a range of topics all at once. I'm only speaking from personal experience on these cumulative finals because I think that I retain more information if I have the incentive to study everything at the end. I don't know if there are any measureable results to prove this.

Carolyn