Improving Learning Outcomes Using an Intelligent Tutoring System

Krishna Vijayaraghavan[1*], and Sasan Ebrahimi[1] and Mst Sunzida Ferdoues[1]

Frequent feedback synchronized with the lesson plan has been shown to improve student learning outcomes in several studies. With increasing enrolments coupled with reductions in education funding, instructors are left with fewer resources to provide frequent feedback or to develop enhanced teaching techniques. Multiple choice questions based assessments, while easier to administer, are considered less reliable than traditional free-form assessments. Intelligent tutoring systems (ITSs) based assessments can improve the speed of assessment while providing immediate feedback to reinforce lectures and free-up resources for instructors. This paper first proposes an ITS design for multi-part problems where all questions are posed at once. Next, the paper evaluates student perception of this new ITS through a survey and a focus group. Following this, the paper evaluates the effectiveness of this ITS on a test score (an indicator of learning outcomes). Finally, the paper lists lessons learned that would be useful to the education community at large. The study indicates that the ITS was received well by the students and that the time taken to complete each attempt of an ITS assignment was comparable to a paper-based assessment. Additionally, the analysis of test scores indicates that the proposed ITS can lead to improvements in student learning. It may be noted that the proposed ITS differs from commercial ITSs by posing all questions at once. However, the pedagogical advantage of offering all questions at once was not evaluated in this study. Further, the proposed ITS was not compared to other ITSs and the observed benefits may not be unique to this implementation. As such, the current findings are locally encouraging and important.

Keywords: ITS, Intelligent tutoring systems, automated assessment system, non-programming assessment, multi-part assessment.

1 Introduction

Frequent feedback has been shown to improve student learning outcomes, particularly when the feedback is integrated and synchronized with the lesson plan (Gibbs & Simpson, 2004). Instant feedback has been shown to have a positive effect on student learning outcomes in as diverse fields as accounting (Mohrwei & Shinham, 2015), clinical nursing for improving concept maps (Wu, Hwang, Milrad, Ke, & Huang, 2012) and computer science (Nutbrown, Higgins, & Beesley, 2016). Nutbrown, Higgins, & Beesley, 2016 note that student were able to identify common mistakes and quickly identify areas of improvement as a result of instant feedback. Currently, instructors are faced with increasing student enrolment (Klemencic & Fried., 2015) coupled with reductions in education funding (Geiger, 2010; Mitchell & Leachman, 2015). This has left instructors with fewer resources to provide frequent feedback or to enhance their teaching techniques. While multiple choice questions (MCQ) are easier to administer, they may not accurately measure student learning (Funk & Dickson, 2011; Stankous, 2016). Intelligent tutoring system (ITS) based assessments can improve the speed of assessment while providing

[1] Mechatronic Systems Engineering, Simon Fraser University, Surrey, BC, Canada V3T 0A3
* Corresponding Author, email: krishna@sfu.ca

immediate feedback to reinforce lectures. Additionally, an ITS would free up time that would be normally allocated to evaluating assignments, as well as eliminate the overhead associated with recording assessment scores. The instructors may invest this time towards developing more interactive and immersive teaching methods or to provide more one-on-one interactions with the students. The automated nature of ITS assessments would aid with data collection in the adoption of outcome-based education.

Published studies provide strong support that ITSs can improve student learning in rudimentary mathematics, particularly at the middle and high school levels (C. R. Beal, Arroyo, Cohen, & Woolf, 2010; Carole R Beal, Walles, Arroyo, & Woolf, 2007; Brusilovsky, 1999; Graesser, Chipman, Haynes, & Olney, 2005; Mitrovic et al., 2009). A good meta-analysis on the effectiveness of the ITSs can be found in Kulik & Fletcher, 2015, who note that ITSs can result in a median improvement of 0.66 standard deviations in the test scores. The research at Pittsburgh Advanced Cognitive Tutor Center (PACT) (J. R. Anderson, Corbett, Koedinger, & Pelletier, 1995; A. Corbett, Koedinger, & Anderson, 1997; K. R. Koedinger, Anderson, Hadley, & Mark, 1997), and its progeny, Carnegie Learning's cognitive tutors, have focused on ITSs for middle and high school students. A very large-scale evaluation of the PACT-ITS indicates that students coached by the ITS outperformed students in a control group by 15% (K. R. Koedinger et al., 1997). A similar study on middle school students suggests that ITSs can lead to an increase in motivation to tackle difficult math problems (Razzaq et al., 2007). Aleven, Mclaren, Sewall, Koedinger, & Mclaren, 2009 have proposed an enhancement to the PACT/cognitive tutors termed "example-tracing tutor", where they compare student behavior against their database and provide step-by-step guidance. Melis et al., 2007 have developed a system called ActiveMath that aims to incorporate more advanced problems such as theorem proofs by using an XML-based representation of mathematical knowledge. In computer engineering, there are several ITS tools such as CourseMarker (Higgins, Hegazy, Symeonidis, & Tsintsifas, 2003), Assyst (Jackson & Usher, 1997), HoGG (Morris, 2003) and Online Judge (Cheang, Kurnia, Lim, & Oon, 2003) for evaluating programing assignments. Ala-Mutka, 2005 and Crow, Luxton-Reilly, & Wuensche, 2018 provide good reviews of several ITS tools for assessing programing skills. They conclude that these systems can provide no feedback and cannot asses all aspects of programming. It may be noted that ITS in programing tend to execute the program (i.e. run the program and examine its outputs) and cannot be extended outside of computer programing.

Outside of rudimentary mathematics and computer programming, ITSs have been utilized in simulating physics experiments (Browne, 2002), support vector machines based evaluation of engineering assignments (Quah, Lim, Budi, & Lua, 2009) and on an integrated testlets for posing MCQ in physics (Shiell & Slepkov, 2015). ITSs have been used in cryptography (AbuEl-Reesh & Abu-Naser, 2018) and for tutoring seismic data interpretation (Ahuja, 2018). Despite showing promising results, ITSs have not been widely adopted to multi-part problems (MPPs) common in post-secondary education. The logical extension of many of the current ITS methods to MPP would result in students being awarded an unfairly low score for mistakes in early parts of the problem (refer to the section "The challenge of assessing multi-part problem"). This would presumably cause student resentment as binary all-or-nothing scoring is known to cause student resentment (Fray, 1989) and there have even been attempts to introduce partial-credit scoring on the multiple-choice exam (Grunert, Raker, Murphy, & Holme, 2013). The integrated testlets discussed earlier (Shiell & Slepkov, 2015) can only pose multi-part MCQs. However, the integrated testlets is better equipped to pose related problems rather than MPP as the solution cannot explicitly depend on solutions to earlier problems. There are a few commercial web-services such as MasteringEngineering (by Pearson) and SaplingLearning that may be applied to MPP. These systems pose one part of the question at a time and use an answer-until-correct approach to "guide" students towards the correct answer before proceeding to the next part of the question. Many MPP particularly in engineering, can involve "design" with multiple correct solutions. Here, subsequent parts of the problem depend on earlier design decisions, and answer-until-correct may be unsuitable. Hence this paper focused on a system that would pose the problem in its entirety rather than individual steps of the problem. This was done in part to encourage the student focus on the "big picture" rather than individual steps and train them on more realistic problems. Further, posing the whole problem would allows the ITS to be easily extended to posing automated exams. Additionally, any grading system will require access to some student information. Privacy policy at our universities (which is in-line with other universities) encourages student personal information to be stored on the university servers (or, with some exceptions, stay within Canada) and strictly prohibits this information being stored with "non-Canadian entities". All commercial products the authors are aware of are hosted outside of Canada. An open source program was chosen so that it can be installed on a computer

managed by the university. Hence this paper proposes the evaluation of a new ITS for MPP that can pose an evaluate the problem as a whole as follow

1. Design of an ITS for MPP to deal with student mistakes in early parts of the assessment.
2. Evaluate how undergraduate students would perceive the ITS.
3. Examine whether the ITS would affect student learning outcomes.

The design of an ITS for MPP is presented below, followed by the study designed to evaluate the ITS. It is postulated that the ITS would allow students' work to be evaluated instantly allowing students to review and re-attempt problems.

The authors would like to note that there are varying definitions of what constitutes an ITS, and the system developed may also be viewed as Computer Aided Instruction (CAI) system as per an alternative interpretation of Kulik & Fletcher, 2015.

2 Implementation of an ITS

2.1 The challenge of assessing multi-part problems (MPP)

To illustrate the challenge with MPP, consider a sample undergraduate design problem consisting of designing a simply supported beam shown in Figure 1. The problem requires the determination of

1. F_A, the reaction force at end A
2. F_B, the reaction force at end B
3. M, the maximum bending moment
4. h, the cross-sectional height (given width w and maximum stress allowable s_{max})
5. I, the area moment of inertia
6. x_{max}, the location of maximum deflection.
7. d_{max}, the maximum deflection (given the modulus E).

The solution to the first three steps of the problem is known to be

$$F_A = F \times (L - a)/L \tag{1}$$

$$F_B = F - F_A \tag{2.i}$$

$$F_B = F \times a/L \tag{2.ii}$$

$$M = F_A \times a \tag{3.i}$$

$$M = F_B \times (L - a) \tag{3.ii}$$

$$M = F \times a \times (L - a)/L \tag{3.iii}$$

$$h \geq [6M/(w \times S_{max})]^{1/2} \tag{4}$$

$$I = \frac{1}{12} wh^3 \tag{5}$$

$$x_{max} = L - [(L-a) \times (L+a)/3]^{1/2} \tag{6}$$

$$d_{max} = \begin{cases} -\left(\left(F_A x_{max} \times (L^2 - (L-a)^2 - x_{max}^2)\right)/(EI)\right), & if\ x_{max} \leq a \\ -\left(\left(F_B (L - x_{max}) \times (L^2 - (L - x_{max})^2 - a^2)\right)/(EI)\right), & if\ x_{max} > a \end{cases} \tag{7}$$

With sample values of $L = 4m,\ F = 5m,\ a = 1m,\ w = 0.01m,\ E = 200GPa$, and $S_{max} = 250MPa$, we find $F_A = 3.75N$ (using 1), $F_B = 1.25N$ (using 2.i or ii), $M = 3.75\ N \cdot m$ (using 3.i or ii or iii), $h \geq 0.003m$. Subsequently for $h = 0.003m$ it is seen that $x_{max} = 1.764m$ and $d_{max} = 0.01863m$.

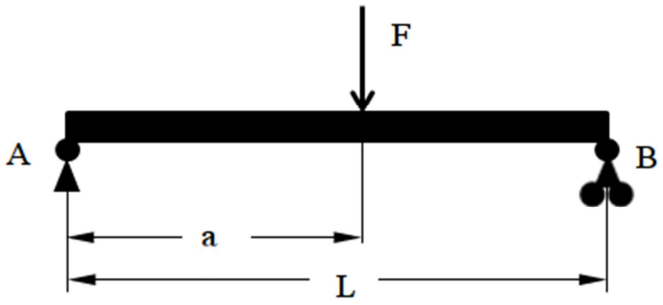


Figure 1: Simply supported beam

Suppose the student incorrectly calculates $F_A = 1.25\ N$ (perhaps using an incorrect formula $F_A = F \times a/L$). The student may calculate $F_B = F - F_A = 3.75N$ (using 2.i), and M = $F_A \times a = 1.25N$ (using 3.i) or $M = F_B \times (L-a) = 11.25\ N \cdot m$ (using 3.ii). This would subsequently cause errors in the values of h and d_{max}. Although the student would have made a mistake only in the very first step, all of the numerical results would be incorrect. If the student receives no credits for the problem, they would likely be frustrated and may be dissuaded from attempting the problem. Further, it should be noted that in a design problem any value of h larger than $0.003m$ is acceptable. The student could have the correct calculation and chosen $h = 0.004m$ and lost points for I and d_{max}. When assignments are evaluated manually, all students receive the same question, and although uncommon, there have been instances of dishonesty. Given the work load, a marker does not have time to rework the problem to incorporate mistakes students make at early stages of the problem, and markers often guestimate the correctness. Occasionally students may end up receiving partial credit even when their approach is entirely wrong and students may have difficulty identifying their mistakes.

2.2 Design of an ITS for MPP

The 1st step of the design process is to identify the different solution paths the students might take as illustrated in Figure 2. This will form the domain model for the ITS. The domain model requires the instructor to have a good grasp of the course material and the problem at hand. The student understanding of the concept (occasionally referred to as "student-model" (VanLehn, 2006) is tracked by the score that the student receives in each course segment and the tutoring strategy consists of highlighting incorrect answers then providing students with the correct numerical solution. This is performed by the inner loop of the ITS. In addition, the inner loop of the ITS is designed to provide each student with randomized values for F, l and a (with $a < l/2$). For the values of $F = 5N$, $L = 4m$, and $a = 1m$, an algorithm as illustrated in Figure 3 may be implemented. Now if the student enters $F_A = 1.25N, F_B = 3.75N$, and calculates the maximum bending moment $M = F_A \times a = 1.25\ N \cdot m$, the program would not provide any marks for F_A and inform them that their answer is incorrect. For each student response, the inner loop reevaluates subsequent steps of the problem. This allows the students to identify mistakes, while simultaneously giving students a full score for F_B and M. The algorithm can be improved by giving partial credits to predictable errors such as incorrect sign. A sample implementation of this algorithm is provided in Appendix A. In this implementation, all potential solutions arising from the different solution paths are included by using "partialcredit(…)". The proposed system uses a rudimentary time based outer loop to pose the assignment after a course module has been completed. A more advanced outer loop would have been more appropriate for a completely self-guided course.

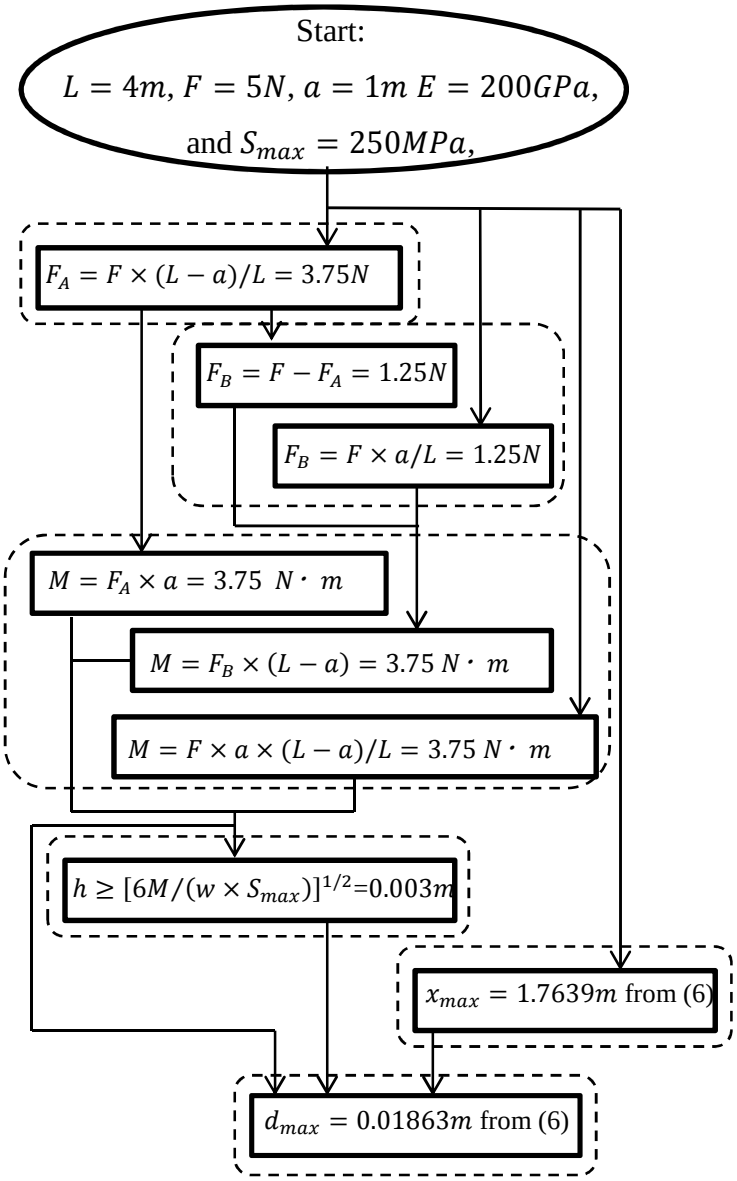


Figure 2: Simply supported beam solution (correct answer)

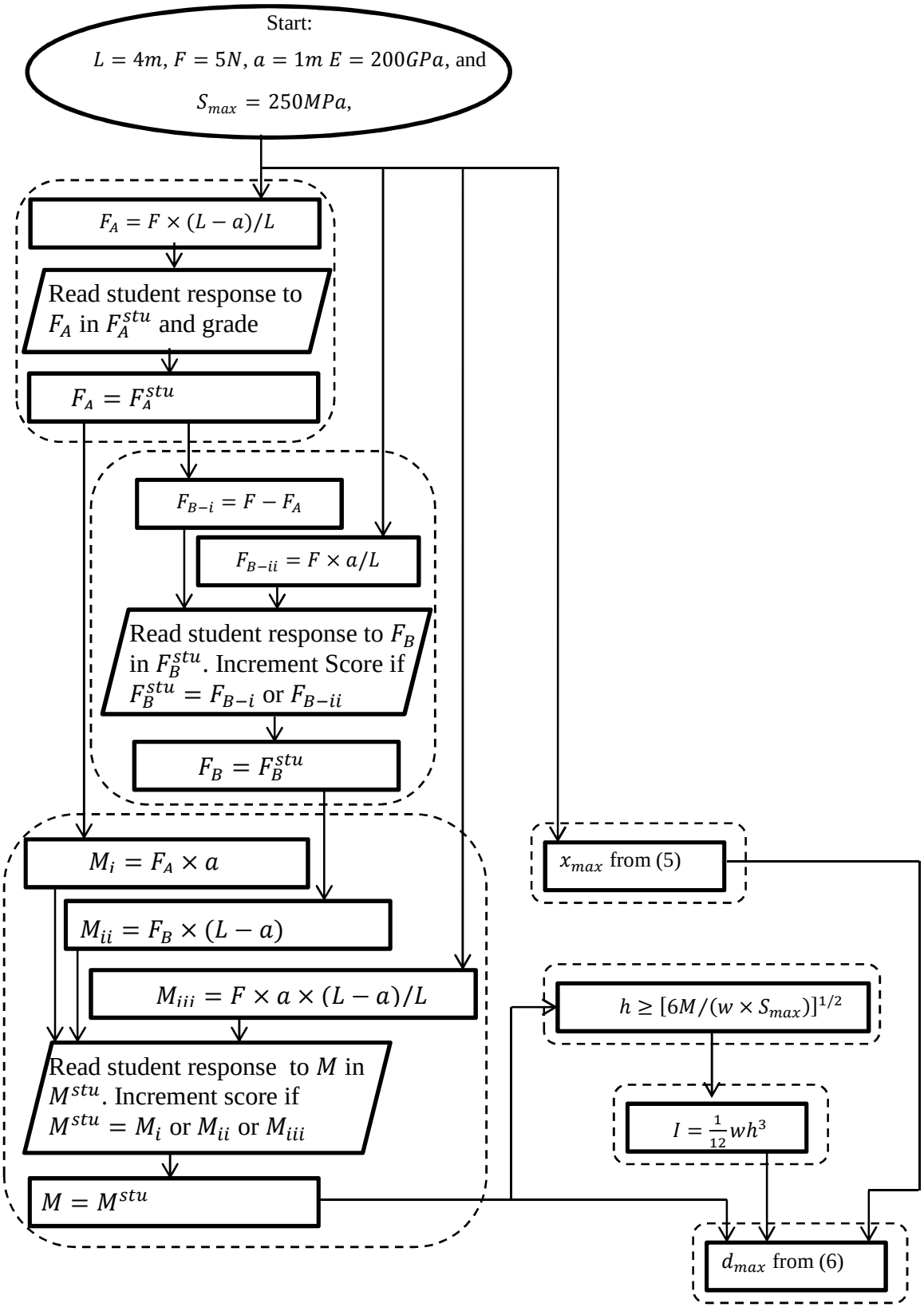


Figure 3: Simply supported beam solution evaluation

3 Study to evaluate of the ITS in MPP

3.1 Implementation of ITS

A senior-level undergraduate course was selected for implementing the ITS. Table 1 provides an overview of the implementation. The course had sequential segments that all student needed to complete. The ITS was only implemented in the first segment (segment 1) of the "new" offering the course. The students were assessed separately on each of the two segments (through test 1 and test 2 respectively). Test 1 in both offerings of the course was 1hr 50min in duration and consisted of numerical problems. Test 2 was changed from a 1 hr explanatory solution based test in the old offering to a ½ hour multiple-choice based test.

The ITS consisted of two assignments that were due at the end of week 3 and week 4. Each ITS assignment had two questions, with each question having between 7 and 17 sub-parts. Students were additionally given ungraded practice assignments identical to the graded assignments. Students could make multiple attempts on the practice assignments with different numerical. Student could choose to utilize the practice assignment either before attempting the graded assignment or reattempt the practice assignment question before the test (or both). Students were given the correct answers to the assignments when they submitted their responses.

Table 1: Overview of the course where ITS was implemented

	Segment 1 (assessment: test1)	Segment 2 (assessment: test2)
NEW offering (87 students)	New segment 1 Utilized ITS ("Treatment") Assessment: $test1_{NEW}$	New segment 2 No ITS Assessment: $test2_{NEW}$
OLD offering (89 students)	Old segment 1 No ITS Assessment: $test1_{OLD}$	Old segment 2 No ITS Assessment: $test2_{OLD}$

The domain model for the ITS consisted of the formulae to solve for loads in trusses, stress in trusses, moments of inertia available in engineering literature (Krenk & Høgsberg, 2013; Mott, 2014). The scoring model (student model), and the inner and outer loops of the ITS system were implemented as discussed in the previous section on a local *IMathAS* server. Details on *IMathAS* can be found in Platz & Niehaus, 2014. *IMathAS* was chosen as it was open-source and can easily be customized to MPP. Students logged into the ITS through an existing learning management system (LMS) called Canvas currently in use at the university. For each

assignment, the numerical values in the problem were randomly generated for each student. The student also had ungraded practice assignments which could be attempted multiple times with different random numerical values. The ITS assignments were mandatory in the new segment 1. 99% of the students completed the ITS assignment receiving an average score of 89% and a median score of 97%. The course also included an optional paper-and-pencil assignment that had been offered in the old offering of the course.

3.2 Evaluation of the ITS

At the end of the new segment 1, students were given an optional anonymous web-survey (refer Appendix B) and were asked to participated in a voluntary focus group (the focus group was moderated by a faculty who was not associated with the course). The students were made aware that their participation and responses in either the survey or the focus group had no bearing on their course grades. Student test score (an indicator of learning outcomes) was used as an objective measure of the effect of the ITS treatment. Since the ITS was implemented as part of a regular course, students could not be split into multiple groups to be given different assignments. However, it is seen from Table 1, that the ITS "treatment" was only applied in the new segment 1. Hence the comparison between $test1_{NEW}$ scores and $test1_{OLD}$ scores would be indicative of the effect of the ITS "treatment". In the absence of a strong control experiment group, the comparisons between $test2_{NEW}$ and $test2_{OLD}$ would serve as a proxy-control in measure any improvement that may be attributable to any difference between the new and the old cohorts of students.

4 Results

4.1 Survey results

Twenty-six out of eighty-seven students responded to the online. The responses have been summarized below (rounded to whole percentage points).

4.1.1 Ease of using the ITS system:

- There was a split opinion about accessing the ITS through the LMS, with roughly 50% of the respondents either agreeing or strongly agreeing the ITS was easily accessible through the LMS and 40% of the respondents disagreeing.
- A significant majority (62%) of the respondents either agreed or strongly agreed that they preferred a dedicated web-access to the ITS.

- On the intuitiveness of the ITS web page, the opinion was split with nearly 35% agreeing or strongly agreeing that the system was intuitive and nearly 42% of the respondents either disagreeing or strongly disagreeing.

4.1.2 Time spent completing assignments

- The majority of the respondents, required the same amount of time for completing the ITS assignments (within 10%) as they would have taken for completing a traditional assignment.

4.1.3 Comparing the course with an ITS to a regular course

- 39% of the respondents (35% agreeing and 4% strongly agreeing) felt that the ITS made doing assignments more enjoyable, while 15% disagreed and 19% strongly disagreed.
- 61% of the respondents either strongly agreed (23%) or agreed (38%) that the instant feedback provided by the ITS increased their motivation to complete the assignments. However, 15% and 4% respectively either disagreed or strongly disagreed with this statement.
- A vast majority (61%), either strongly agreed (15%) or agreed (46%) that the ITS made review easier.

4.1.4 Perception of learning outcomes

- Only 23% of the respondents (4% strongly agreed and 19% agreed) indicated that they retained more information through the ITS, with 42% of the respondents were neutral, and 31% either disagreeing (19%) or strongly disagreeing (12%).
- 35% of respondents either strongly agreed (8%) or agreed (27%) that the ITS improved their overall understanding of the course.
- 35% of respondents either strongly agreed (8%) or agreed (27%) that the ITS helped or would help with their overall scores in the course
- 61% of respondents reattempted the assignments

4.1.5 Course preference

- 46% of the respondents were neutral about preferring a course with an ITS. There was nearly an equal number of positive and negative responses to preferring such a course.

- The effect of the ITS on the likelihood of the student registering for an elective was inconclusive.

4.1.6 What students found most valuable with ITS

The terms “instant feedback” followed by “reattempts” and “different values” (which would refer to the students’ ability to reattempt the problems with different numerical values) was the most commonly used words to describe the most valuable feature of ITS. One respondent stated that the most valuable feature of the ITS as being “the possibility for instant feedback, the ease of access, and the ease of submission”. The respondent adds that “These things made it more likely for me to attempt the assignments at all”. A full examination of the responses indicates that students found the “instant feedback” to be useful and they also liked the ability to reattempt the questions with different values. A Wordle^{TM} of the responses is presented in Figure 4, where “instant-feedback” features prominently, followed by “reattempt” and “different values” (the term “assignment” may be ignored).

(Note: Words implying the same intent such as, “instant answers” and “instant feedback”, are lumped together)

Figure 4: Wordle^{TM} of “What I found most valuable about the automated assignments (ITS) ”

4.1.8 Suggestions to improve the automated assignment system

Although the program was coded to propagate the student-entered solution one respondent felt that “mistake early on meant the error compounded”. Some students felt that the ITS was an extra step and one respondent felt that the system was “another hoops to jump through”. There was a general concern about the difficulty in logging into the system. There were concerns that reattempting a question completely resets the problem and students had to go through the entire problem again to determine where they went wrong. Respondents also suggested improving the quality of feedback by including a “mechanism” for detecting “students making predictable

mistakes". There was concern about students sharing solution templates in either Excel or Matlab format.

4.2 Focus group results

Nine students participated in the focus group that was led by a faculty member with the Teaching and Learning Center (TLC) at Simon Fraser University (SFU). From the focus group, the following categories emerge.

Clarity and ease of use:

The focus group had mixed opinions about the clarity and ease of use. They found the system "weird" and "wonky", as the system had "unexpected functionality". The students also had mixed opinions about the email instruction and different students had varying degrees of success getting the system to work on different browsers. Some in the group also found the user-interface to be "counterintuitive" and to have "odd-colors". After submitting the questions the system "said 'you submitted' ", but the students were unsure if the whole question or only a part had been submitted. The main concern with the ease of use arose from the presence of spelling and mathematical mistakes. However, students did not raise any major issues with replacing paper assignment with an ITS. The students did not state that the ITS graded them unfairly. However, they did state that the ITS may be unfair if used as an exam.

Multiple attempts on the problems:

The students were disgruntled that the questions would "reset" for each reattempt and that they had to "re-work" the problem in its entirety to spot their mistakes. The group felt that the system "did not let students see their work". The opinion regarding the usefulness of the immediate feedback was mixed. The group expressed concern that the "only feedback was whether or not the answer was right" and that the feedback lacked explanation. However the group "very much liked getting immediate feedback". Nonetheless, the group would like "more qualitative feedback [sic]" (better quality feedback). Everyone in the group reworked the practice problem "many times" before completing and submitting the assignment.

Effect on learning outcomes:

The group "thought it helped improve their understanding" of the course. The group also felt that the ITS would be a "useful tool for assignments and learning" if the ITS is "programmed" properly. Further everyone "re-did their work" on practice problems.

4.3 Analysis of student test scores

Figure 5 shows the distribution of $test1_{NEW}$ (test after ITS was implemented) and $test1_{OLD}$ scores in segment 1 (please refer Table 1 for the overview of the implementation). Students obtained a mean score of 58% (standard deviation 16%) in $test1_{OLD}$ (without ITS) and a mean score of 88% (standard deviation 16%) in $test1_{NEW}$ (with ITS). This indicates two standard deviations improved in test scores in segment 1 as a result of the ITS. Figure 6 shows the distribution of test scores for segment 2. In this proxy-control test, the mean decreased from 73% (standard deviation 14%) in $test2_{OLD}$ to 45% (standard deviation 16%) in $test2_{NEW}$. While this may indicate that the new groups of students were not inherently smarter than the old group, there is insufficient evidence to draw a definite conclusion as the drop in score may have resulted from test2 being changed from a problem-based assessment to a multiple choice based assessment. Figure 7 plots the regression between $test2_{NEW}$ and $test1_{NEW}$, and between $test1_{OLD}$ and $test2_{OLD}$. The linear regression between that $test1_{OLD}$ and $test2_{OLD}$ has a slope 0.452 and R2 = 0.1729, while $test1_{NEW}$ are clustered between 80%-100% indicating an across the board improvement in student test scores. Students who scored in the lower 50% percentile of $test2_{OLD}$ had an average score of 52.5% in $test1_{OLD}$, while Students who scored in the lower 50% percentile of $test2_{NEW}$ had an average score of 87.1% in $test1_{NEW}$. Students who scored in the upper 50% percentile of $test2_{OLD}$ had an average score of 61.2% in $test1_{OLD}$, while Students who scored in the upper 50% percentile of $test2_{NEW}$ had an average score of 89.7% in $test1_{NEW}$.

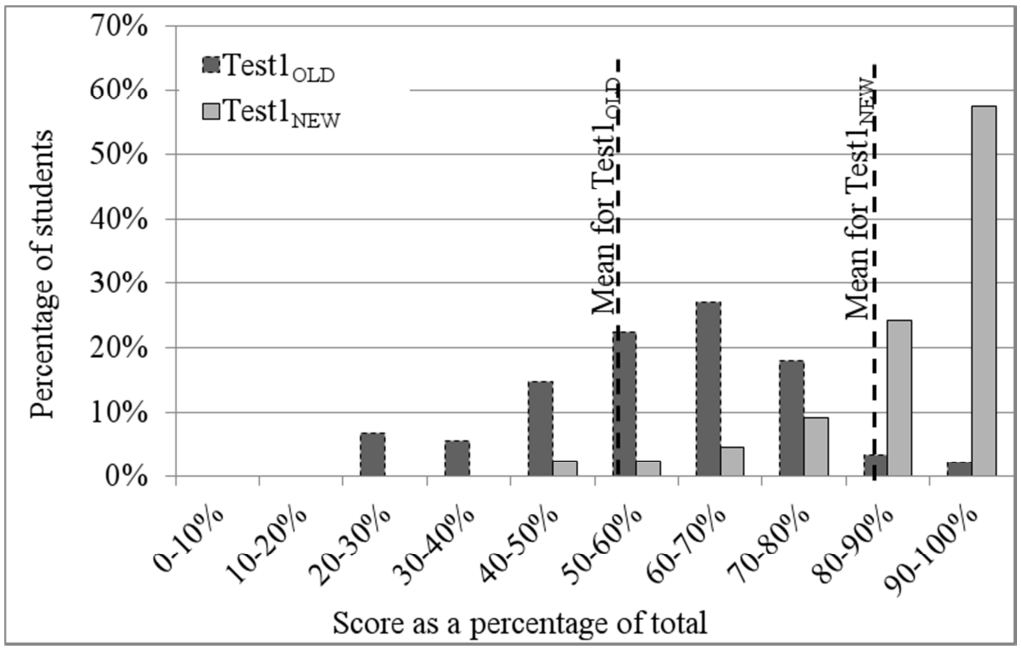


Figure 5: Distribution of student test 1 scores for two offerings of the course

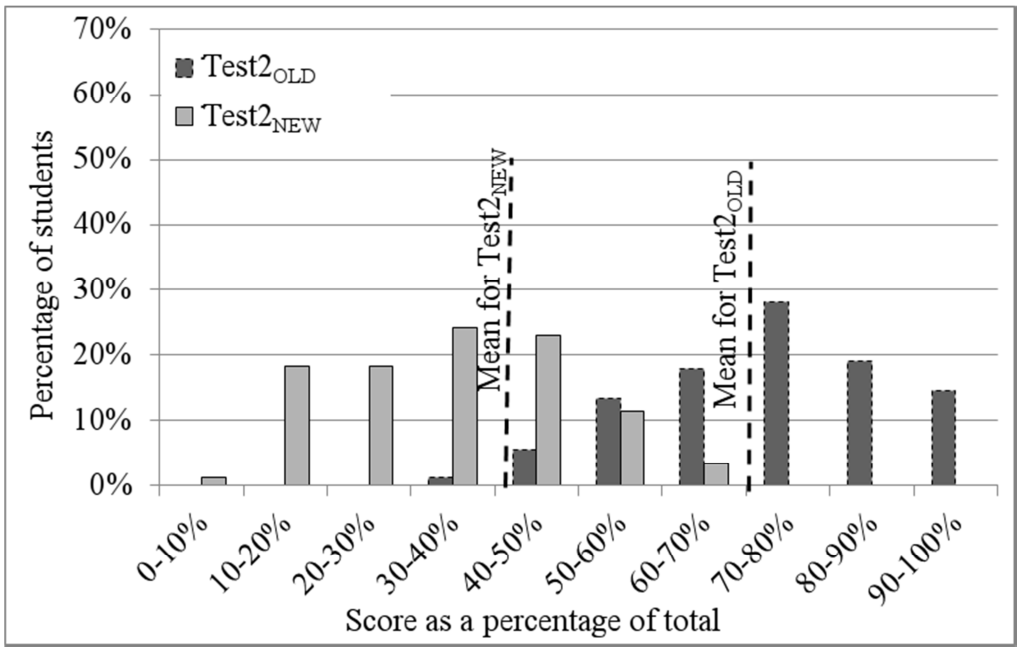


Figure 6: Distribution of student test 2 scores for two offerings of the course

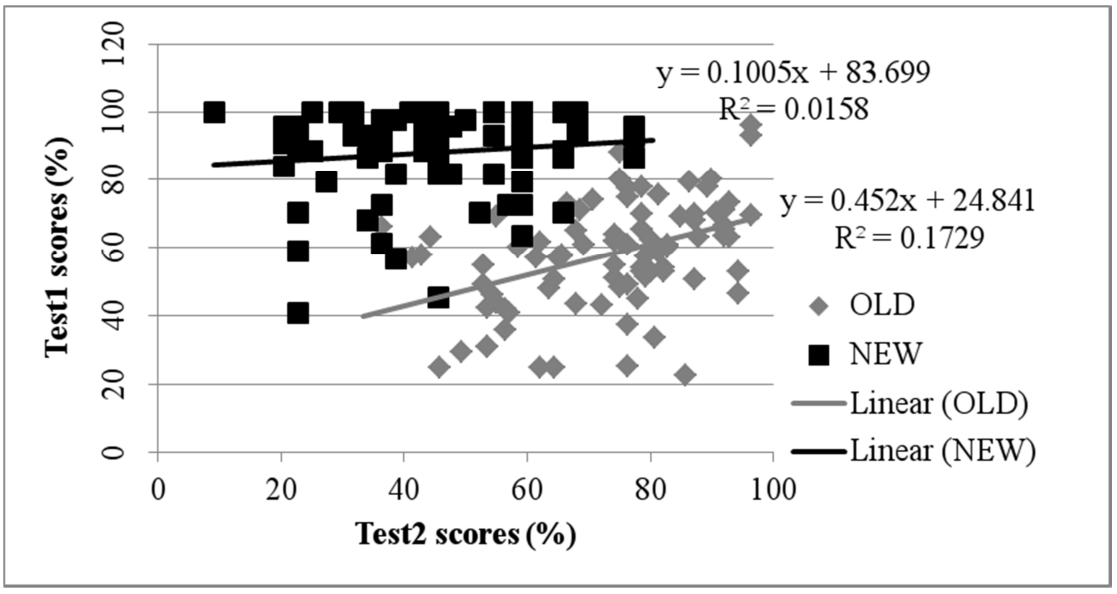


Figure 7: Plot of test 1 scores as a function of test 2 scores for two offerings of the course

The Kolmogorov-Smirnov test for normality of the test scores shown in Table 2 revealed that $test1_{NEW}$ were not normally distributed, while the Levene Mean larger than an alpha-value of 0.05 indicates an absence of Homogeneity of variance in test 1. A single factor ANOVA between $test1_{OLD}$ and $test1_{NEW}$ (Table 3) indicates a significant change in test2 scores (with F-critical

3.895 for an alpha level of 5%). It may be noted that a single factor ANOVA comparing $test2_{OLD}$ and $test2_{NEW}$ in Table 3 also indicates a significant change in test2 scores as well. ANACOVA of the $test1_{OLD}$ and $test1_{NEW}$ scores (using the $test2_{OLD}$ and $test2_{NEW}$ scores as a covariate) resulted in an F-value 177.53, a p-value <0.001 (Table 4). However, these results should be interpreted with some caution as $test1_{NEW}$ is not normally distributed.

Table 2: Test for normality and homogeneity of variance in test score

Test	Kolmogorov-Smirnov (K-S) test for normality			Homogeneity of variance
	K-S statistic	Critical K-S statistic	Are scores normally dist.	Levene Mean
$Test1_{OLD}$	0.11	0.144	Yes	0.3849
$Test1_{NEW}$	0.165	0.146	NO	
$Test2_{OLD}$	0.11	0.144	Yes	0.03131
$Test2_{NEW}$	0.13	0.146	Yes	

Table 3: ANOVA of test scores

Test	F-value	P-value
$Test1_{OLD}$	199.192	1.22×10^{-30}
$Test1_{NEW}$		
$Test2_{OLD}$	150.168	2.73×10^{-25}
$Test2_{NEW}$		

Table 4: ANACOVA of test scores

Test	Covariate	F-value	P-value
$Test1_{OLD}$	$Test2_{OLD}$	177.53	2.51×10^{-28}
$Test1_{NEW}$	$Test2_{NEW}$		

5 Analysis of results

5.1 "How would undergraduate students perceive the ITS"

There was some ambivalence both in the web-survey and in the focus group discussion. Some of the ambivalence may be attributed to the use of the new system and some programming

errors from initial deployment. The clarity and ease of use could be improved by further increasing the amount of testing and by hiring additional programmers and TAs during initial deployment. With the exception of one survey respondent, students did not raise any major issues with using an ITS in the place of a paper assignment. Additionally, the students required the same amount of time they would have needed to complete a traditional assignment. The student perception may be further improved by incorporating an FAQ page on the ITS.

5.2 "How would the ITS affect student learning"

While the survey revealed that the students did not feel the ITS made assignments more enjoyable, it nonetheless increased their motivation to complete the assignment and the ITS provided students with the opportunity for additional practice. The students did not perceive that the ITS helped with the overall understanding of the course. The students were ambivalent about registering for a course solely because it used an ITS. The test scores from Figure 5 indicate that the student learning outcome may be improved by using an ITS. More significantly, with the use of the ITS, roughly 57% of the students obtained a score between 90-100% and 24% of the students got a score in the range of 80-90%. A single factor ANOVA between $test1_{OLD}$ and $test1_{NEW}$ revealed which revealed an F-value larger than F-critical and a low P-value indicating a significant difference between the scores after the ITS-treatment. Additionally, there was a significant improvement in test scores and a narrowing of the gap between the test scores of students who scored in the lower 50% and the upper 50% on the reference test. Part of this improvement may be attributed to the introduction of graded assignment (from ungraded assignments), however, the earlier literature suggests that the introduction of graded (as opposed to ungraded) assignments only improves the performance of freshman students (Grove & Wasserman, 2006) and inexperienced college students (Geide-Stevenson, 2009). This indicates the entire student group was benefiting from the ITS and the learning gap between students may be reduced by using an ITS.

6 Discussion of results

The results indicate that the students had a positive experience with the ITS. Students perceived that they spent the same amount of time on each attempt of the ITS as they would have otherwise spent on a traditional assignment. It appears that the students were motivated to re-attempt the problems related to the course material. From the authors' experience, although some students may review assignment questions, they do not normally attempt to re-solve assignment

problems (particularly if the numerical values are unchanged). Although it was not explicitly stated, it may be inferred that the immediate feedback motivated the students to reattempt the problem. Improving the quality of feedback requires full integration of the course with the ITS with lecture notes being linked to the ITS. While this is beyond the scope of the current study, this should be investigated in future.

7 Lessons learned

7.1 What worked

The ITS did not appear revolutionary to students and there was little resistance to the adoption of the ITS. The students did like the overall experience and the ITS web-interface did not overly burden students. The ITS reworked subsequent stages of the problems using student answers. This feature was well received and with the exception of one feedback. Students also liked the ability to rework the problems and indeed almost all students reworked the problems.

7.2 Sources of problems (what did not work)

Occasionally, students mistyped answers into the answer box. The ITS subsequently reworked the problem, using the mistyped answers. Fortunately, there were only a few instances of mistypes and in all instances, students' scores were manually updated by the instructor. A more robust scheme is required for the next implementation of the ITS. The solutions displayed were based on a combination of correct values and student entered values, which resulted in some confusion when a student attempted to rework the problems. It may be better to only display the answer based on the correct solutions. The ITS was designed to reset the questions for each student attempt. Students were frustrated that they had to rework the entire problem to identify mistakes. Using LMS credentials to login into the ITS had mixed results. Hence, the ITS should also be tested across multiple browsers and the login made seamless. Despite the best of intentions, the complexity of the ITS systems resulted in unforeseen errors. With the immediate feedback, students are less tolerant of such errors. Some students were frustrated that they were being treated as beta-testers. Students also anticipate a higher quality of feedback that would help them identify the errors which would require closely working with past students to identify useful feedback. There were a few minor comments regarding the color scheme and layout used in the ITS.

8 Conclusion

The study developed the design of an ITS system for MPP common in the undergraduate curriculum. Each time a student entered a value, the ITS would re-work subsequent stages of the problem using the value entered by the student, thereby mitigating the effects of early mistakes. Encouraging students to act upon feedback (from corrected assignments) has been challenging in the best of situations. The ITS gave the opportunity for students to reattempt problems without any additional workload on the instructor and almost all students reattempted the problems. Hence, the ITS may improve students' understanding of the course material. The ITS can also help instructors identify mistakes in the students' understanding of concepts, as well as common numerical errors students', may make.

The immediate feedback implies any errors in the solution become evident immediately and the students are less tolerant of such errors. There is scope for improving the ITS by including better quality feedback perhaps through a video tutorial. Overall the study indicates that ITS may offer significant benefits to students in non-programming engineering problems. The ITS may improve student overall understanding of the course and result in better test scores. Equally important is the fact that an ITS in assessment can significantly reduce instructor load without adversely affecting the student experience and learning outcomes. The results of this paper can be expanded to education in other faculties such as mathematics, physics, and chemistry with assignments based on formulae. Further, the ITS-assessments would lend themselves to data collection and data mining for the purposes of implementing outcome-based education.

It may be noted that the proposed ITS differs from commercial ITSs by posing all questions at once. However, the pedagogical advantage of offering all questions at once was not evaluated in this study. Further, the proposed ITS was not compared to other ITSs and the observed benefits may not be unique to this implementation. As such, the current findings are locally encouraging and important.

9 Acknowledgment

The authors would like to acknowledge funding from the Institute for the Study of Teaching and Learning Centre in the Disciplines (ISTLD) at Simon Fraser University (grant G0125). The authors would also like to acknowledge invaluable support and guidance by Vivian Neal, Dr. Cheryl Amundsen, Dr. Angela McLean and Dr. Gregory Hum with the Teaching and

Learning at SFU. The authors would also like to thank the anonymous reviewers for their comments that greatly helped improve the quality of the paper.

10 Exemption from ethics board review

The ITS system was developed and implemented as part of curriculum enhancement at the school of Mechatronic Systems Engineering. Hence, the evaluation falls within the provisions of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2 2014 Article 2.5) and was thus exempt from ethics board review. An exemption letter from Office of Research Ethics at Simon Fraser University was provided to the journal.

References

AbuEl-Reesh, J. Y., & Abu-Naser, S. S. (2018). An intelligent tutoring system for learning classical cryptography algorithms (ccaits). *International Journal of Academic and Applied Research*, *2*(2), 1-11.

Ahuja, N. J. (2018). Characterization of human knowledge for intelligent tutoring. In K. Saeed, N. Chaki, B. Pati, S. Bakshi, & D. P. Mohapatra (Eds.), *Progress in Advanced Computing and Intelligent Engineering* (pp. 363-373). Singapore: Springer Singapore.

Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments. *Computer Science Education*, *15*(2), 83-102. https://doi.org/10.1080/08993400500150747

Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: example-tracing tutors. *International Journal of Artificial Intelligence in Education,*, *19*(2), 105-154.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: lessons learned. *Journal of the Learning Sciences*, *4*(2), 167-207. https://doi.org/10.1207/s15327809jls0402_2

Beal, C. R., Arroyo, I., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of animalwatch: an intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning*, *9*(1), 64-77. Retrieved from http://secsem.cs.arizona.edu/projects/focal/ergalics/files/BealArroyoCohenWoolf_2010.pdf

Beal, Carole R, Walles, R., Arroyo, I., & Woolf, B. P. (2007). On-line tutoring for math achievement testing: a controlled evaluation. *Journal of Interactive Online Learning*, *6*(1), 43-55.

Browne, R. F. (2002). Automated tutorial and assignment assessment. *Journal of Educational Technology & Society*, *5*(1), 119-123. Retrieved from http://www.jstor.org/stable/jeductechsoci.5.1.119

Brusilovsky, P. (1999). Adaptive and intelligent technologies for web-based education. In *Special Issue on Intelligent Systems and Teleteaching, Künstliche Intelligenz* (Vol. 4, pp. 19-25). Retrieved from http://www2.sis.pitt.edu/~peterb/papers/KI-review.html

Cheang, B., Kurnia, A., Lim, A., & Oon, W.-C. (2003). On automated grading of programming assignments in an academic institution. *Computers & Education*, *41*(2), 121-131. https://doi.org/10.1016/S0360-1315(03)00030-7

Corbett, A., Koedinger, K., & Anderson, J. (1997). Intelligent tutoring system. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (pp. 849-874). Elsevier Science.

Crow, T., Luxton-Reilly, A., & Wuensche, B. (2018). Intelligent tutoring systems for programming education: a systematic review. In *20th Australasian Computing Education Conference* (pp. 53-62). Brisbane, Australia. https://doi.org/10.1145/3160489.3160492

Fray, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79-96.

Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology*, *38*(4), 273-277. https://doi.org/10.1177/0098628311421329

Geide-Stevenson, D. (2009). Does collecting and grading homework assignments impact student achievement in an introductory economics course ? *Journal of Economics & Economic Education Research*, *10*(3), 3-15.

Geiger, R. L. (2010). Impact of the financial crisis on higher education in the united states. *International Higher Education, 59*, 9-11. https://doi.org/10.6017/ihe.2010.59.8486

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students ' learning. *Learning and Teaching in Higher Education*, *1*(1), 3-31.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). Autotutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612-618. https://doi.org/10.1109/TE.2005.856149

Grove, W. A., & Wasserman, T. (2006). Incentives and student learning: a natural experiment with economics problem sets. *American Economic Review*, *96*(2), 447-452. https://doi.org/10.1257/000282806777212224

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: development of a rubric for rating partial credit. *Journal of Chemical Education*, *90*(10), 1310–1315. https://doi.org/10.1021/ed400247d

Higgins, C., Hegazy, T., Symeonidis, P., & Tsintsifas, A. (2003). The coursemarker cba system: improvements over ceilidh. *Education and Information Technologies*, *8*(3), 287-304.

Jackson, D., & Usher, M. (1997). Grading student programs using assyst. In *Twenty-eighth SIGCSE technical symposium on Computer science education* (Vol. 29, pp. 335-339). New York, NY. https://doi.org/10.1145/268085.268210

Klemencic, M., & Fried., J. (2015). Demographic challenges and future of the higher education. *International Higher Education, 47*.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Krenk, S., & Høgsberg, J. (2013). *Statics and mechanics of structures*. London, UK: Springer.

Kulik, J. A., & Fletcher, J. D. (2015). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, *XX*(X), 1-37. https://doi.org/10.3102/0034654315581420

Melis, E., Andres, E., Budenbender, J., Frischauf, A., Libbrecht, P., Pollet, M., … Goduadze, G. (2007). Activemath : a generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education*, *12*, 385-407.

Mitchell, M., & Leachman, M. (2015). *Years of cuts threaten to put college out of reach for more students* (pp. 1-26). Retrieved from https://www.luminafoundation.org/files/resources/year-of-cuts-threaten-to-put-college-out-of-reach.pdf

Mitrovic, A., Martin, B., Suraweera, P., Zakharov, K., Mcguigan, N., & Zealand, N. (2009). Aspire : an authoring system and deployment environment for constraint-based tutors. *International Journal of Artificial Intelligence in Education*, *19*, 155-188.

Mohrwei, L. C., & Shinham, K. M. (2015). Enhancing students' learning: instant feedback cards. *American Journal Of Business Education*, *8*(1), 63-70.

Morris, D. S. (2003). Automatic grading of student's programming assignments: an interactive process and suite of programs. In *33rd ASEE/IEEE Frontiers in Education Conference* (pp. 1-6).

Mott, R. L. (2014). *Machine elements in mechanical design* (5th ed.). Upper Saddle River, NJ: Pearson.

Nutbrown, S., Higgins, C., & Beesley, S. (2016). Measuring the impact of high quality instant feedback on learning. *Practitioner Research In Higher Education*, *10*(1), 130-139. Retrieved from http://ojs.cumbria.ac.uk/index.php/prhe/article/view/318

Platz, M., & Niehaus, E. (2014). Imathas & automated assessment of mathematical proof. *Beiträge Zum Mathematikunterricht*, *2014*, 915-918. https://doi.org/10.17877/DE290R-15649

Quah, J. T.-S., Lim, L.-R., Budi, H., & Lua, K.-T. (2009). Towards automated assessment of engineering assignments. In *International Joint Conference on Neural Networks* (pp. 2588-2595).

Razzaq, L., Feng, M., Heffernan, N. T., Koedinger, K. R., Junker, B., Nuzzo-jones, G., … Walonoski, J. A. (2007). A web-based authoring tool for intelligent tutors : blending assessment and instructional assistance. *Studies in Computational Intelligence*, *44*, 23-49.

Shiell, R. C., & Slepkov, A. D. (2015). Integrated testlets: a new form of expert-student collaborative testing. *Collected Essasy on Learning and Teaching*, *8*, 201-210. https://doi.org/10.22329/celt.v8i0.4244

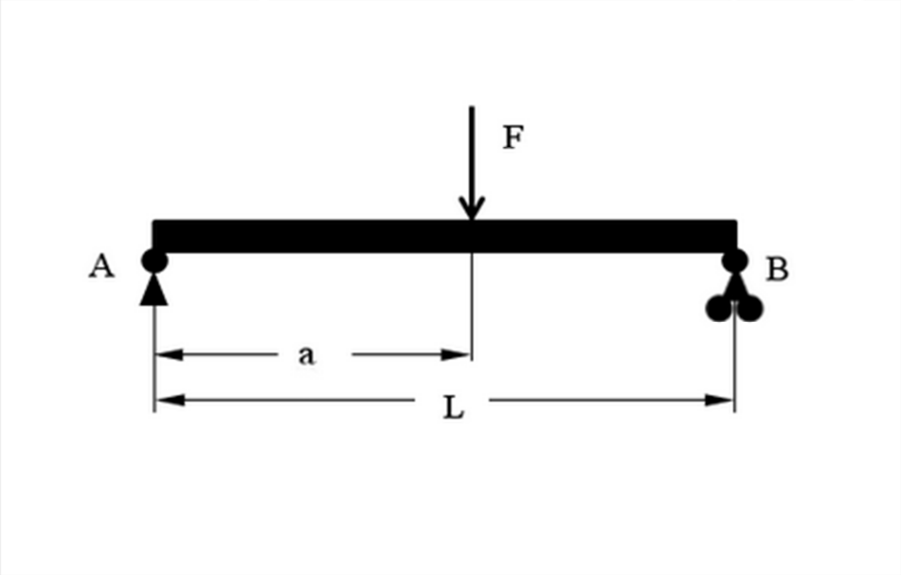
Stankous, N. V. (2016). Constructive response vs . multiple-choice tests in math : american experience and discussion (review). *European Scientific Journal*, *7881*(May), 308-316.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227-265.

Wu, P., Hwang, G., Milrad, M., Ke, H., & Huang, Y. (2012). An innovative concept map approach for improving students' learning performance with an instant feedback mechanism. *British Journal of Educational Technology*, *43*(2), 217-232. https://doi.org/10.1111/j.1467-8535.2010.01167.x

Appendix A: Sample fair ITS program

A simply supported beam AB is shown below. Assume that the length of the beam is 4 m, the magnitude of the downward force is 5 N and that the force is applied at a distance of 1 m from end A. Assume that the Young's Modulus of the beam is 200 GPa and yeild strength of 250 MPa.



Calculate

The reaction force at end A: [ ] Preview N (wt 15%)

The reaction force at end B: [ ] Preview N (wt 15%)

The Maximum bending moment: [ ] Preview N-m (wt 40%)

The minimum thickness of the beam if its width 0.01 m [ ] Preview m (wt 5%)

The area moment of inertia of the above beam [ ] Preview m^4 (wt 10%)

The distance of the point of maximum deflection from the end A [ ] Preview m (wt 5%)

The maximum deflection of the beam [ ] Preview m (wt 10%)

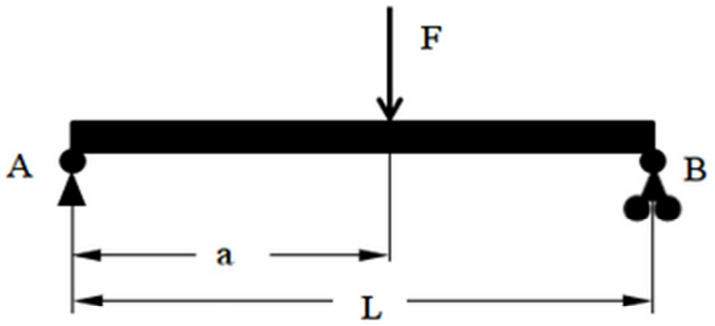
Figure 8: Student interface with ITS

Score on last attempt: 1.5 out of 10 (parts: ✔ 1.5/1.5, ✖ 0/1.5, ✖ 0/4, ✖ 0/0.5, ✖ 0/1, ✖ 0/0.5, ✖ 0/1)
Score in gradebook: 1.5 out of 10 (parts: ✔ 1.5/1.5, ✖ 0/1.5, ✖ 0/4, ✖ 0/0.5, ✖ 0/1, ✖ 0/0.5, ✖ 0/1)

No attempts remain on this problem.

This question, with your last answer and correct answer, is displayed below

A simply supported beam AB is shown below. Assume that the length of the beam is 4 m, the magnitude of the downward force is 5 N and that the force is applied at a distance of 1 m from end A. Assume that the Young's Modulus of the beam is 200 GPa and yeild strength of 250 MPa.



Calculate
The reaction force at end A: 3.75 Preview N (wt 15%)
The reaction force at end B: Preview N (wt 15%)
The Maximum bending moment: Preview N-m (wt 40%)
The minimum thickness of the beam if its width 0.01 m Preview m (wt 5%)
The area moment of inertia of the above beam Preview m^4 (wt 10%)
The distance of the point of maximum deflection from the end A Preview m (wt 5%)
The maximum deflection of the beam Preview m (wt 10%)

Answer: 3.75
Answer: 1.25
Answer: 3.75
Answer: 0.003
Answer: 2.25E-11
Answer: 1.7639320225002
Answer: 0.018633899812498

Figure 9: Student interface with ITS upon submission

```
1.|  $anstypes=array("calculated","calculated","calculated")
2.|
3.|  $L = rand(3,8);
4.|  $a = rand(0.1*$L,0.9*$L);
5.|  $F = rand(1,10);
6.|
7.|  $FA=($L-$a)/$L*$F;
8.|
9.|  $answer[0] = $FA;
10.|  $answweights[0]=0.25; $reltolerance[0]=0.01;
11.|  $partialcredit[0]=array(-$FA,0.5);
12.|  $FA=$stuanswers[0] if( ($stuanswers[$thisq][0]!=null) &&
      ($stuanswers[$thisq][0]!=0) );
13.|
14.|  $FB_i = $F-$FA; $FB=$FB_i;
15.|  $FB_ii = $a/$L*$F;
16.|  $answer[1] = $FB_i;
17.|  $answweights[1]=0.25; $reltolerance[1]=0.01;
18.|  $partialcredit[1]=array(-$FB_i,0.5,-$FB_ii,1, -$FB_ii,0.5,   );
19.|  $FB =$stuanswers[1] if( ($stuanswers[$thisq][1]!=null) &&
      ($stuanswers[$thisq][1]!=0) );
20.|  $M_i=$a*$FA; $M=$M_i;
21.|  $M_ii=($L-$a)*$FB;
22.|  $M_iii = $a*($L-$a)/$L*$F;
23.|  $answer[2] = $M_i;
24.|  $partialcredit[2]=array(-$M_i , 0.5,$M_ii,1, -$M_ii,0.5, -$M_iii,1, $M_iii,0.5);
25.|  $answweights[2]=0.5; $reltolerance[2]=0.01;
```

Appendix B: Survey template

Evaluation of IMathAS (ITS)

Agreement of participation

Effectiveness of Automated Assignment on course outcomes

As part of the course improvement, an automated assignment system (ITS) known as IMathAS was used in to generate all assignments for one part of a course you took. You were required to complete all the assignments for this part of the course using this automated system. This survey aims to evaluate the effectiveness of the system.

Your participation will involve completing this short survey. Your participation is voluntary; you have the right to decline. If you choose to participate, you may withdraw from the evaluations of the automated system at any time with no consequences to your education or your grade in the course.

The findings of this evaluation will be shared with the school of Mechatronic Systems Engineering, members of the SFU community and may be shared with educators beyond SFU through presentations, written reports and articles. Please feel free to contact me krishna@sfu.ca if you have any questions.

This survey is completely anonymous and no identifying information is being collected. At the end of the survey, you will be provided with the link to be entered into the raffle. By completing this survey you agree to participate in this evaluation.

Q1 . Ease of using the system

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
It was easy to access the IMathAS (ITS) via Canvas (the learning management system) :	○	○	○	○	○
I would prefer a dedicated webpage to access the IMathAS (ITS) :	○	○	○	○	○
The webpage and interface of the automated system was intuitive to use :	○	○	○	○	○

Q2 . Time spent completing assignments

	#1 took over 25% less time than #2	#1 took 10-25% less time than #2	About same time (within 10%)	#1 took over 10-25% more time than #2	#1 took over 25% more time than #2
How much time did the IMathAS (ITS) (#1) take to complete, compared to traditional assignments (#2) :	○	○	○	○	○

Q3 . Comparing course with automated assignment (IMathAS) to a regular course

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree

Compared to traditional assignments, automated assignment (ITS) made doing the assignments more enjoyable :	○	○	○	○	○
The ability to review the automated assignment (ITS) instantly (compared to paper assignments that are returned after 1-2 weeks) increased my motivation to complete the assignments:	○	○	○	○	○
The automated assignment (ITS) made review easier:	○	○	○	○	○
I retained more information using the automated assignment (ITS) :	○	○	○	○	○
Automated assignment (ITS) helped to improve my overall understanding of the course material :	○	○	○	○	○
The automated assignment (ITS) helped/would help me attain better score in the exams :	○	○	○	○	○
I reattempt the assignments with different numerical values as a practice for the exam (or I would in future courses with automated assignment (ITS)) :	○	○	○	○	○

Q4 . Course preference

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I would prefer a course with automated assignment (ITS) over a traditional course :	○	○	○	○	○
The use of the automated assignment (ITS) would make me more likely to register for an elective course compared to a traditional course :	○	○	○	○	○

Q5 . What I found most valuable about the automated assignment (ITS)

Q6 . Suggestions to improve the automated assignment system (ITS)