Advances in Engineering Education

SUMMER 2015

Engineering Student Self-Assessment Through Confidence-Based Scoring

GIGI YUEN-REED
IBM T. J. Watson Research Center
Yorktown Heights, NY

AND

KYLE B. REED
University of South Florida
Tampa, FL

ABSTRACT

A vital aspect of an answer is the confidence that goes along with it. Misstating the level of confidence one has in the answer can have devastating outcomes. However, confidence assessment is rarely emphasized during typical engineering education. The confidence-based scoring method described in this study encourages students to both think about their answers in a different way and to evaluate their confidence in the answer. Each answer is scored based on whether the answer is right or wrong and whether the student is confident or not in that answer. Students generally appreciated the educational value as it made them more self-aware of their understanding. Overall, students were able to accurately assess whether their answer was right or wrong 77% of the time. Average self-assessment generally improved over time, but the degree of improvement varies based on student segments. The method also benefits instructors by indicating the topics that students tend to be less certain of, even if the students are getting the right answers, and identifies students that are either over or under confident.

Key Words: confidence-based scoring; certainty-based markings; self-efficacy; assessment

INTRODUCTION

An important skill for students is to recognize how confident they are in their stated answers to questions. Even if students get the right answers, they may not be sure it is right and may get similar questions wrong. Confidence is vital in engineering jobs and in graduate education where the confidence students have in their solutions can be as important as the solutions themselves. However, confidence in one's answer or design is not typically emphasized during undergraduate engineering courses. Information is becoming easier to quickly obtain from digital sources like Wikipedia, internet search engines, and online literatures, but the reliability of that information is not necessarily better. Thus, the confidence in a solution based upon that information is important for one to evaluate and report (Gardner-Medwin 2008).

The literature has discussed many frameworks for describing problem solving, metacognition about the answer, and how the resulting solution can lead to an increased understanding (Schoenfeld 1992). One way of thinking about a solution consists of four aspects: the information used, the method applied, the stated answer, and the confidence one has in that answer. Undergraduate education typically focuses on two of those: methods and answers. In fact, typical grading schemes tend to encourage students to guess or "just put something down" in hopes that they will get partial credit, which does not encourage a deeper understanding of their missing knowledge (Gardner-Medwin 1995). Confidence-based scoring (CBS) aims to address the last aspect – providing the incentives and mechanism for students to assess and state their confidence in their answers (Gardner-Medwin 2006). CBS is designed to combine their answer with their confidence so they are encouraged to further examine their answers and more seriously evaluate their confidence level. When the students recognize that they do not know the answer, they get some credit for correctly stating their lack of knowledge or understanding. When the students get the correct answer but are not sure, they do not get the full credit due to the uncertainty. This method of self-assessment encourages students to evaluate their abilities throughout a course and enables them to become more aware of this important aspect of decision-making (Gardner-Medwin and Gahan, 2003). This method also intends to encourage students to practice accurate self-assessment of their ability so that they can leverage that insight to succeed in their professional life.

Two ways that confidence is commonly referred to include the common catchphrases "fake it until you make it" or "trusting your life with it". "Fake it until you make it" describes the idea of appearing confident in oneself despite potential shortcomings or uncertainties and encourages continual efforts such that one eventually masters the skill. Adams (2011) summarizes several research studies that demonstrate how faking a positive behavior can lead to emotional and health benefits. Cuddy et al. (2015) demonstrate that maintaining a pose that exudes power and confidence for two minutes changes the hormone levels to more closely match those of people who are more often in leadership roles; in other words, faking the physical characteristics can actually affect the behavior. Although not the focus of this paper, the practice of appearing confident in one's ability, whether or not actually being confident, is encouraged in domains like entrepreneurship and research/sales proposals. The practice tends to promote optimism and persistence to allow individuals to achieve the desired results and often introduces opportunities that would not be available if one appears hesitant.

Nevertheless, not all professions or situations are suited for the practice of "fake it until you make it." In engineering design, practitioners must conduct quality checks and often employ safety factors in their design to ensure a "trust-your-life-with-it confidence". For example, if an engineer said "This engine might be ready for flights," one would be hesitant to use it on a commercial airline. However, if the competent engineer signed off and stated "This engine is definitely ready for use in passenger planes," the engine would be shipped off and installed. In many ways, the practice of being confident in one's answers is essential for consumer safety and the engineering professions' reputation. While different levels of confidence tend to suit different professional situations or decision types, they are all rooted in one's ability to assess his/her knowledge in the subject matter.

This paper presents and analyzes a confidence assessment method based on a binary choice (either confident or not confident), which is a simplification of the more complex assessment methods in the literature (Hassmen & Hunt 1994; Gardner-Medwin 1995). Moreover, our work applies CBS to a range of question types including multiple choice and open-ended questions. Previously used CBS methods are typically applied to multiple choice or true/false questions (Barr and Burke, 2013).

For clarity of understanding and conciseness, the following definitions are given:

- *Right* – getting the one true answer
- *Wrong* – any answer that is not right
- *Correctness* – adjective describing whether the answer is right or wrong
- *Confident* – student has a high level of belief that their answer is right; the student's answer is considered confident unless they explicitly state otherwise (i.e., default answer)
- *Not confident* – student is unsure of their answer; student marked the not-confident box
- *Accurate assessment* – correctly judging the correctness of their answer; i.e., selecting confident when they are right and not confident when they are wrong

BACKGROUND

The idea that a person's confidence is an important part of knowledge can be traced back to Confucius (circa 500BC) and Aristotle (circa 300BC) as described by Hunt (2003). More recently, confidence has been studied using frameworks that generally have the goal of encouraging a deeper understanding of the material (Hunt 2003; Heron and Lerpiniere 2013), increasing reflection and justification of one's answers (Cisar, Cisar, and Pinter 2009), and as an assessment tool (Hevner 1932; Gardner-Medwin and Gahan 2003). Self-assessment studies are often described in terms of confidence and/or self-efficacy (Bandura 1977; Gecas 1989; Pajares 1996). Bandura (1997) defines confidence and self-efficacy as follows: "confidence is a nondescript term that refers to strength of belief but does not necessarily specify what the certainty is about" and "self-efficacy refers to belief in one's agentive capabilities, that one can produce given levels of attainment." The study presented in this paper focuses on an individual's judgment about whether he/she is certain of each individual answer or not.

Survey-Based Methods

Several survey-based methods have been used to examine confidence and self-efficacy. Bandura (2006) used several scales and surveys for determining self-efficacy on a range of tasks including exercise, driving, eating habits, and others. He measures the degree of self-efficacy on general tasks using a scale of 0 to 100. Carberry et al. (2010) incorporated aspects of Bandura's (2006) surveys to conduct a study on a self-concept interest specifically related to how engineering students perceive their abilities by asking students to rate their self-efficacy. They found the survey was able to identify engineering students' self-efficacy on general design problems, motivation, and anxiety. Other surveys using different scales have also been used to ask about self-efficacy in fields such as writing (Pajares and Johnson 1994), engineering (Kolar and Carberry 2013), and music education (Jeanneret 1997). These studies use surveys that are not directly tied to specific questions, but are related more generally to their knowledge and ability in a field.

Several additional survey studies provide insights regarding specifically how engineering students perceive their skills. Parsons et al. (2009) found that first year engineering students who were better mathematically qualified were generally more confident and successful in mathematics. A study by Fantz et al. (2011) found that students who had hobbies related to engineering and students who had pre-engineering classes had significantly higher self-efficacy measures than students without these interests or extra classes in their first year. Another study showed that individuals try harder to solve a given problem when they see someone they perceive as similarly competent solve the same problem (Brown and Inouye 1978). Ponton et al. (2001) suggest that professors can enhance a student's self-efficacy by developing skills, peer interaction, encouraging students, and explaining coping strategies, all of which are important for practicing engineers.

Impact of Confidence on Learning

Information Reference Testing (Bruno 1993) formalized the interlinked relationship of developing one's knowledge and confidence in the learning process. The study aimed to simultaneously identify a person's knowledge and confidence in that knowledge. This results in two metrics where mastery is defined as being both confident and knowledgeable, which generally leads to smart actions. Low confidence and low knowledge lead to an uninformed person that hinders taking action. High confidence and low knowledge describes a misinformed person that leads to mistakes. Low confidence and high knowledge describes a person with doubts and leads to hesitation. This framework is most closely aligned with the method tested and discussed in this paper since our method uses two states each for confidence and correctness, where knowledge and correctness are related, but differ in their scope.

The retention of new knowledge is correlated with the confidence in the learned material. Hunt (2003) found that people will only remember 25% of material after a week when they stated they were "not sure at all" of their answer. Those that stated "extremely sure" retained 91% of their learned knowledge after a week. This finding suggests that helping students both learn and become more confident in new material can help them retain the knowledge better. However, negative biases in self-evaluations can affect satisfaction with learning compared to those with a positive self-evaluation bias, even though the actual performance between both groups is the same (Narciss et al. 2011). Learning to accurately assess one's knowledge is, thus, important for both the learning process and later in one's career when applying the learned knowledge.

Emergence of Confidence-Based Scoring

One of the earliest academic papers studying the use of confidence-based scoring hypothesized that the reliability of grading would be improved by incorporating confidence into the student's answer since a correct answer on a multiple choice question always has the relatively high chance of being a guess (Hevner 1932). Hevner designed several grading schemes that score students' answers based on different combinations of the number of right and wrong answers. She found that the scoring scheme using correct answers that were weighted based on the student's stated confidence had the highest correlation to six other measures of knowledge including training and talent.

Over the years, researchers have continued to study various confidence-based assessment methods. The CBS methods studied tend to evaluate multiple levels of confidence. One of the repeatedly-studied methods for CBS includes three levels of confidence with a correct or wrong answer (Gardner-Medwin & Curtin 2007; Gardner-Medwin 2013). Other versions have used five levels of confidence (Hassmen & Hunt 1994; Khan et al. 2001) and one used 11 levels (Petr 2000). Dissatisfaction was high among students in the study with 11 levels (Petr 2000), but this is not the common reaction to CBS based on larger studies with fewer levels of confidence (Gardner-Medwin and Curtin 2007). The different impressions of the method may be based at least partially upon the cost (e.g., time and added stress during an assessment) balanced with the benefit the students understand they may get from enhancing their self-assessment skills.

A slightly different implementation separates the question about the knowledge from the determination of their confidence (Rosewell 2011). As opposed to the typical CBS method, students in

Rosewell's study stated their confidence on the problem before they could see the multiple-choice answers to select from. Like many of the studies that use CBS, this method is implemented on a computer. This revised method helps to transform a multiple-choice question into an open-ended question since the students must formulate their answer and determine their confidence prior to making the multiple choice selection. In comparison, this paper applies confidence on both open-ended and multiple choice questions, and the Conclusions and Future Work section proposes some additional methods to incorporate confidence into more complex questions that benefit from grading with partial credit, which is common in higher level engineering courses.

STUDY DETAILS AND DATA COLLECTION METHOD

Confidence-Based Scoring Method Used in This Study

Confidence-based scoring (CBS) uses both the correctness of the answer as well as the student's selection of "confident" or "not confident" to determine the grade for each question. Table 1 shows the points awarded for each of the four combinations of confidence and correctness in this study. The two axes of correctness and confidence level is similar to the concept presented as Confidence-Based Learning (Bruno 1993) and Confidence-Based Assessment (Gardner-Medwin, 2013), but this version is the simplest version containing only two possibilities for each.

The four quadrants of Table 1 can be thought of in the following way for each specific question answered. Having a right answer but lacking confidence, is beneficial, but not perfectly so. This would be similar to asking a colleague to double check your work, which should increase the confidence in that work, but still requires the use of someone else's time. The worst possible case is to be confident of a wrong answer, so zero points are assigned. The best case is to have the right answer and be confident in it. The final situation is wrong and not confident in the answer. In some ways, this is similar to simply saying: "I do not know," which is the perfect answer when a person truly does not know; *making up an answer is a very bad habit that is encouraged under typical grading schemes*. In this case, the student would get some credit for knowing that they do not know and that they should ask for help.

	Confident	Not confident
Right	5	4
Wrong	0	2

Table 1. Points awarded for each combination of correctness and confidence.

The scores assigned to each of the four categories listed in Table 1 are chosen to encourage students to consider that "not confident" is an acceptable answer and to admit they do not know when they are unsure. This method offers more incentive for students to select "not confident" since they gain 2 points for a wrong answer, but only lose 1 point for a right answer. For example, if students are guessing between two answers, they will get a higher score on average by selecting not confident. This grading method reinforces the idea that stating an answer that one is not certain of is the best alternative to having a correct answer as long as it is stated with the condition of not being confident. The specific score values could be changed to encourage students taking courses outside their major to select confident more often since students tend to underestimate their belief in knowledge outside of their discipline (Knight and Smith 2010).

Although confidence is not usually a clear binary choice, this method aims to encourage students to seriously think about their knowledge of the subject while minimizing the additional cognitive burden. Since students are generally motivated to try to get as many points as possible and confidence is an important part of their grade, students are motivated to put some thought into their choice. Some students may always check the "not confident" box, so they are guaranteed a minimum score of 40%, but they are also limiting their maximum score to 80%. Some students may never check the box marked "not confident" if they are always confident in their work. Ideally, students will think about how sure they are of each individual answer and respond accordingly with different confidence markings for different questions.

Participants

The study included 137 senior undergraduate students majoring in mechanical engineering at the University of South Florida in Tampa. The students participated over three separate Fall semesters of the same course taught by Dr. Reed (one of the authors). None of the students included in this study were repeating this class nor had they previously taken a similar class. They may have had previous exposure to some of the concepts, but we did not consider their prior knowledge of this material in this study.

The participants consisted of 14 females and 123 males. On average, 7.8% of the questions were not answered due to student absences, so the following data includes a total of 1,769 problems. Dr. Reed or the class TA graded the quizzes and each student verified the grading of their own quizzes when returned. Each question has one clear right answer.

The University of South Florida's Institutional Review Board approved this study and the data collection procedure. All students had the study explained to them and they were freely given the option of participating in this study or not participating in this study. All students were graded the same in either case, but only those students that agreed to be included in the study and signed the consent form, hence explicitly participating in the study, were included in this analysis. The signed consent forms were handled by an independent third party during the IRB consenting process and were only returned to the authors after the grading period ended to avoid any possible bias related to participating or not.

Data Collection

The CBS method described in above was integrated into every quiz (but not on any exams) in a senior-level engineering Mechanical Controls course, except for the first quiz. The first quiz was intentionally excluded to get the students accustomed to the all-or-nothing grading (i.e., no partial credit). This first quiz was used to make the point that there are clearly right and wrong answers and if you do not get the right answer, it is wrong. Either the engine will power the plane and make it fly or it will fail. The remainder of the quizzes had some questions with a checkbox to indicate that they are "not confident" in their answer. By default (not marking the box), they are confident. Future versions of this study could include separate checkboxes for "confident" and "not confident" to avoid any possible ambiguity about the student's choice. However, all students included in this study explicitly agreed to participate with full understanding of the procedure and grading technique. The problems on quizzes with the "not confident" checkbox are scored out of five total points as shown in Table 1. Students were shown Table 1 at least two days prior to the first quiz with CBS and again during each quiz, so they were aware of the scoring method before and during each quiz. No specific instructions were given to the students before or after any of the quizzes other than how the scoring works. The only feedback provided to the students was the quizzes themselves that indicated their score that was based on the confidence and correctness of their answers. An anonymous questionnaire was distributed to students during the last week of class (after all quizzes had been taken, graded, and returned) asking for the student's opinions about confidence-based scoring; the qualitative survey results are discussed in the Student Feedback Section below.

Although the quiz questions used in this study are specific to this course, any questions that have a clear right and wrong answer can use this method. Some possible methods to incorporate partial credit are described in Conclusions and Future Work section. For reference, the general topics of the questions used in this research are shown in Table 2 and three example questions are shown in the appendix. The quiz questions stayed the same each year with slight changes (for example, changes in numeric values) between each year to reduce the chance of "sharing" from students who already took the course. The consistency of each question ensures a similar level of difficulty between the semesters. No students took the course twice during this time. A two-way ANOVA with interaction effects was performed to evaluate if the fourteen questions were perceived similarly across the years.

Question #	Quiz #	Concept of the problem
1	2	Block diagram reduction
2	3	Laplace transform
3	3	Final Value Theorem
4	3	Block diagram reduction
5	4	Order of a system from a Bode plot
6	4	System response from a step input
7	4	System parameters from a transfer function
8	4	System parameters from a Bode plot
9	5	Gain margin from a bode plot
10	5	Phase margins from a bode plot
11	5	Gain margin from a bode plot
12	5	Phase margins from a bode plot
13	6	Binary logic truth table
14	6	Ladder logic multiple choice question

Table 2. Summary of quiz questions that used confidence-based scoring (examples of specific questions in the appendix).

The dependent variable was the score on each question and the independent variables were class year (1-3) and the problem number (1-14). The results show that there was no statistically significant interaction effect between the problems and the years (F(1727,26) = 1.2, p = .21).

RESULTS AND DISCUSSION

The three classes studied had very similar overall quiz performance. Table 3 provides the overall information about the three classes and quiz results. The average size of each class was 45.7 students. On average, 62% of the answers were right and 63% of the answers were reported as "confident".

	Class 1	Class 2	Class 3	Overall
Students	44	47	46	137
Questions answered	546	631	592	1,769
Right answers	60% (324)	64% (400)	63% (374)	62% (1,098)
Confident answers	63% (338)	65% (402)	62% (370)	63% (1,110)

Table 3. Overall statistics of the quiz results.

The total number of questions answered in each class varies due to the slightly different number of students and absences throughout the semester. A two-way ANOVA with interaction effects was performed with scores on each quiz question as the dependent variable and independent variables of class year (1–3) and the problem number (1–14). The results show that there is not a statistically significant difference between the three different classes (F(1727,2) = 1.6, p = .19). The results also show that there is a statistically significant difference between the performance on the problems (F(1727,13) = 14.6, p < .00001) throughout the semester, which is to be expected with different material and question types. A separate one-way ANOVA showed that there is a statistically significant difference between students (F(1619,136) = 3.43, p < .00001). The variability between students is much higher than the variability between years. Given the similarity of the quiz questions each year and the lack of statistical difference between the performance over the three classes, the subsequent discussions in this paper are based on the combined data from all three classes.

Accurate Assessment of the Answer

The data shows that confidence tends to correlate with correctness ($R^2 = 0.45$). Table 4 provides the breakdown of answers by correctness and confidence. While 82% of the right answers are associated with a confident selection, only 32% of wrong answers are associated with a confident selection. Similarly, while 68% of the wrong answers are associated with a not-confident selection, only 18% of right answers are associated with a not-confident selection. Students tend to select "confident" when they have the right answer, and select "not confident" when they have the wrong answer. More often than not, students accurately assess the answer so they obtain the maximum score possible (i.e., optimizing their confidence selection). The maximum obtainable score for each question is calculated by assuming a confident selection for correct answers (5 points) and a not-confident selection for wrong answers (2 points). Similarly, the minimum score is calculated by assuming a confident selection for wrong answers (0 points) and a not-confident selection for correct answers (4 points).

Students tend to be better at optimizing their confidence selection (i.e., more accurate) with the right answers than wrong answers. The data shows that accurate confidence selections are more common with right answers (82%) than wrong answers (68%). There are several possible explanations

Correctness	Confident	Not Confident	Row total
Right	82% (897)	18% (201)	100% (1,098)
Wrong	32% (213)	68% (458)	100% (671)

Table 4. Breakdown of correctness of an answer and confidence selection.

to such behavior. One explanation is that students are inherently confident in their abilities and are unaware of their mistakes; it is unlikely that the whole class would have a perfect assessment of their answers. Furthermore, it is reasonable to interpret that students with right answers tend to be more self-aware and are able to better optimize their confidence choice than those with wrong answers. Regardless of the underlying explanation, this finding reinforces the importance of working with underperforming students to better assess their capabilities so that they can focus more on areas that need improvement and also to enable them to seek needed help. Here, underperforming can mean the traditional view of low scores, but in the context of CBS, underperforming can also mean low accuracy in self-assessment.

Students accurately assessed their answers 77% of the time, but could generally improve their score if they were more accurate in their assessments for every question. While CBS encourages students to select confident with right answers and not confident with wrong answers, students need to have an accurate assessment of their knowledge of the course material in order to maximize their scores. Perfect assessment is not expected since students are learning the material. The observed sub-optimal confidence selection suggests some room for improvement. We searched through all of the answers to find those with an accurate assessment and those with an inaccurate assessment. Accurate assessment is defined as selecting confident when right or not confident when wrong; inaccurate assessment is defined as not confident when right or confident when wrong. Accurate assessment maximizes the scores for a given correctness. By analyzing the 1,769 answers, we identified that 414 answers (23%) could have received a better score if students could more accurately assess their abilities. Keeping the answer the same, but assuming a perfectly accurate assessment, the average score would increase from 3.2 to 3.6 (out of 5 points), corresponding to an approximately 10% improvement. On the other hand, if the students were perfectly inaccurate in assessing their ability (i.e., selecting not confident when right and confident when wrong), the average score would decrease from 3.2 to 2.3, corresponding to an approximately 30% reduction in the average score. Figure 1 shows the actual scores with bars that represent the maximum and minimum of scores possible assuming their answer does not change and only their confidence response is changed.

Under or Over Confident

To gain deeper insights into individual students' behavior, we examined the distribution of confidence and correctness. Figure 2 is a scatter plot where the size of each data point represents the number of students. The y-axis represents the percentage of right answers, and the x-axis represents the percentage of answers that confident is stated. The diagonal line represents the students that were accurately assessing their answers; on average, 19% of the students were on the diagonal line. The further away the students are from the diagonal, the worse they

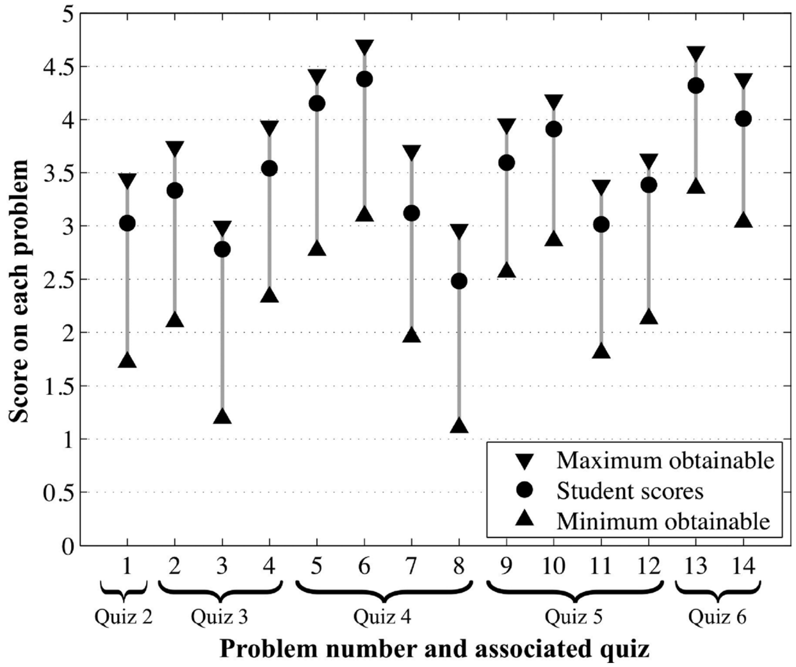


Figure 1. The average student scores on each problem are shown with the range of scores that are possible given that students were perfectly accurate (maximum obtainable; only confident–right and not-confident–wrong) or perfectly inaccurate (minimum obtainable; only not-confident–right and confident-wrong) in assessing whether their answer was right or wrong.

are in accurately assessing their abilities. The figure highlights two groups of students that are not typically noticed, but have behaviors that can be detrimental to engineering decisions. The students on the top left typically get the right answer, but are under confident in the answer. This group will not assert their answer and may be convinced to go along with another solution posed by a more confident individual. On average throughout the semester, 35% of students were under confident. The second group, shown on the bottom right, is composed of students that tend to state confident more often than they get the right answer. This group is more hazardous than the first since they would be likely to implement wrong solutions and may convince others to go along with them. On average throughout the semester, 46% of the students were overly confident. As opposed to helping only those students who do not perform well on the quizzes, these students represent two groups that should also receive guidance to help them better assess their work.

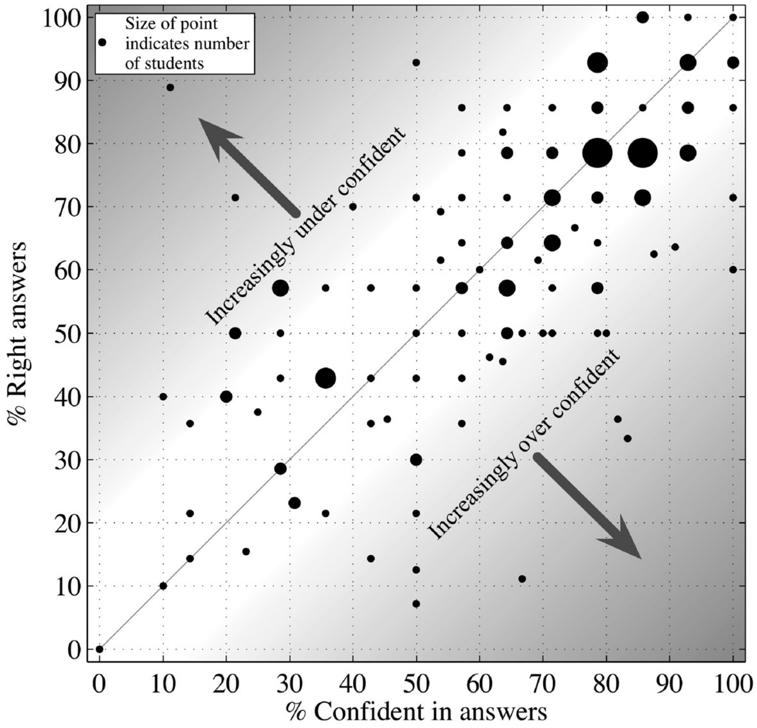


***Figure 2**. The relationship between students' having the right answer vs. being confident based on the percentage of the problems they answered. Individuals to the bottom right of the diagonal line are overly confident (46%) and individuals to the top left are under confident (35%). The 19% of students on the line are, on average, accurately assessing their answers to the questions.*

Gender Influences

Studies in the literature have demonstrated that gender is a factor related to self-perception of engineering abilities (Marra et al. 2009; Riegle-Crumb and Moore 2013). The percentage of females in mechanical engineering tends to be low in general. As such, although our study only includes 14 females, it is important to include our data in the literature. Table 5 shows the quiz statistics divided by gender. Our data shows a slight indication that female students are less confident than male students and generally do not get the right answer as often, but there is no evidence of a gender-difference in their ability to choose an appropriate confidence level. The data shown here suggests that correctness and confidence should be jointly studied (i.e., accurate self-assessment) when examining gender in engineering and other fields.

	Male	Female
Students	90% (123)	10% (14)
Questions answered	90% (1,584)	10% (185)
Right	63%	54%
Confident	64%	56%
Accurate Confidence Selection	77%	76%

Table 5. Overall statistics of the quiz results by gender.

Self-Assessment Over Time

Overall, students are accurately choosing their confidence 77% of the time and demonstrated a slight increase of three percentage points between the first two quizzes and the last three. Looking only at the group average performance occludes interesting aspects about the changes throughout the semester, thus the following analysis divides the students into subgroups.

To more clearly observe assessment accuracy trends throughout the semester, we segmented the students based on their initial accuracy assessment. Segmenting the students this way allows for an analysis of how subgroups within the entire sample change over time; the group average only had a slight change throughout the semester. Specifically, the segmentation is based on their ability to accurately assess their answers on quizzes 2 and 3 (i.e., the warm-up period), which consists of four quiz questions that incorporate this CBS method. The groups are based on the percentage of accurate assessments. All students were able to accurately assess at least one of their answers during the warm up period, thus there are inherently four groups based on a segmentation using the four questions on quizzes 2 and 3. One segment included four students who accurately assessed their answers on only 25% of the first four questions; this group is called the "initially 25%" group. The "initially 50%" group contained 28 students that accurately assessed their answers half the time. The "initially 75%" group contained 41 students and the "initially 100%" group contained 41 students. The following analysis excluded the 23 students who did not complete one or more of the problems during the warm up period since it was unclear to which group they should be assigned.

Although the four groups started at very different self-assessment accuracy levels, they converged to nearly the same ability by the end of the course. As illustrated in Figure 3, students that were initially 100% accurate demonstrated deterioration in self-assessment accuracy over time. The deterioration could be partially explained by the increase in the difficulty of the material over the duration of the course, which would make it more difficult for students to accurately self-assess. However, despite the increasing course difficulty over time, students that were initially 25% and 50% accurate demonstrate a trend of improved self-assessment. A one-way Kruskal-Wallis non-parametric analysis of variance shows that there was a statistically significant difference between

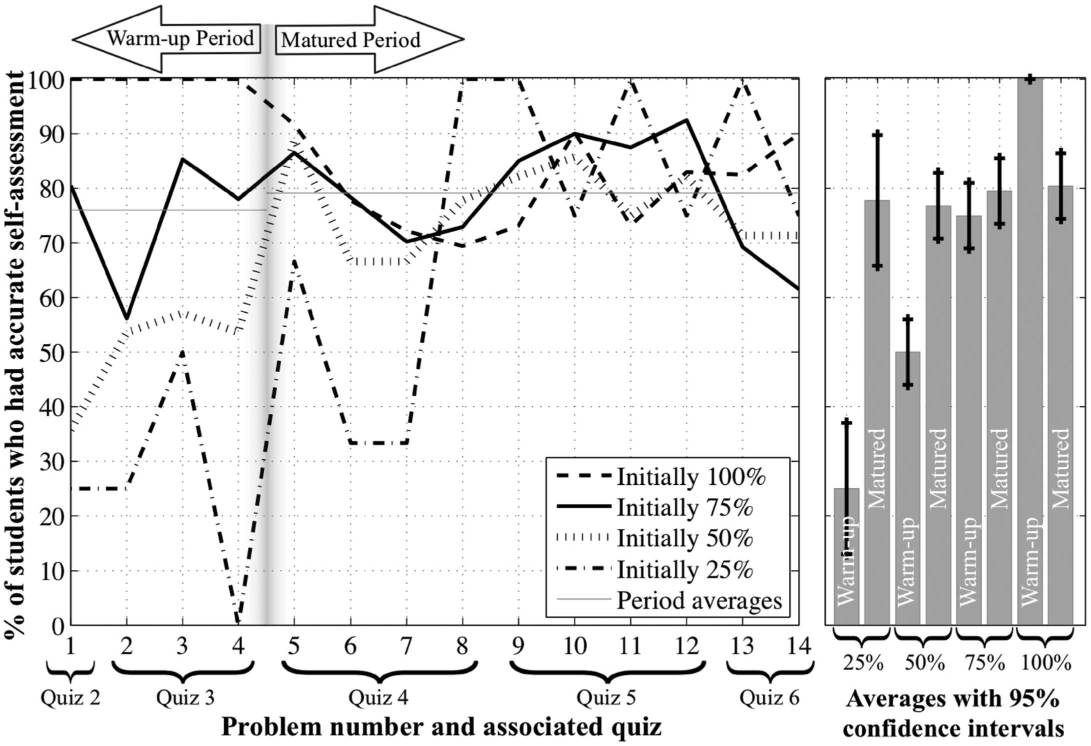


Figure 3. The students' self-assessment ability throughout the course changes. The students were divided into four segments based on their accuracy on the first two quizzes (warm-up period). All four groups converged to accurately assessing approximately 77% of their answers over time. The bar graph shows the averages for each segment in both periods with 95% confidence intervals.

the four segment groups crossed with the warm-up/matured periods (F(1534,7) = 19.7, p < .0001). A post-hoc test using Tukey's Honestly Significant Difference test showed a significant difference between the warm-up periods of the 25%, 50%, and 100% group to all other periods and groups, but there were no differences among the four groups during the matured period. This analysis suggests that, while students may have different initial abilities, self-assessment ability can be improved and the CBS method can be one of the tools to facilitate the learning process.

While this analysis shows that self-assessment can be improved, the reason why students from all groups with different initial accuracies converge to approximately 77% accurate is not fully understood. It is expected that the accuracy would converge to less than 100% since the students are learning the material for the first time and it is unrealistic to have near-perfect assessment time after time. In fact, we observe that self-assessment accuracy fluctuates since the content of the questions change and students adapt to each specific question. However, further studies are needed to uncover the reasons of the convergence and the converged value. The convergence may be driven by the subject area, the composition of the student body, the design of the CBS, and other factors.

Student Feedback

Students filled out an anonymous survey at the end of the semester regarding the class as a whole with several questions specifically asking about the CBS method. The survey was conducted in class with an 81% response rate. Two open-ended questions were asked about the method. First question was "What are your general thoughts about the confidence-based grading?" and the second question was "Did having the option to answer 'not confident' help you to evaluate your understanding of the material?" After the researchers evaluated the responses, it was clear that the student responses could be placed in one of two categories: (1) did not like any aspect of the method and (2) appreciated the value of the method or explicitly stated they liked it. The vast majority of responses were clear stating "I HATE IT", "Yes" (to the second question), "I liked it", or some variation of "no, but it made me think different about the questions." The 12 responses that were not clear or did not have answers were not included in this analysis.

Of the responses that had clear answers, 72% of the students appreciated the method. Not all of these students necessarily liked the method, but indicated explicit appreciation of its value. This is particularly apparent with those who had worked in industry and/or had an internship. The other 28% of the students disliked the method. These students prefer partial credit since they can show that they are on the right path and they believe their grade would have been higher. Below are several representative statements from the responses to the two open-ended questions:

- "I like the way it is done cause there is a reward for knowing the subject and for giving it a try at the same time."
- "It seems like it helps those who are wrong more than those who are right."
- "As a student I find it annoying, but since I had an internship last summer I can appreciate the principle behind it."
- "I am too nervous in making a mental error and get zero points when saying I am confident."
- "The 'not confident' answer made me second guess myself at times but was an indicator that I didn't fully grasp some of the subject matter."
- "It's definitely a double edged sword. However, it has gotten me to look at each question with a new perspective."
- "[The method] stimulates extra effort to look at the material."
- "We as graduating engineers should not only know the material taught, but also know how well we know it."

APPLICATIONS

Confidence-based scoring can be leveraged in several domains. In the classroom, it can promote self-assessment, encourage students to understand their shortcomings, and urge students to seek help accordingly. It can also help support instructors by allowing them to quickly identify students who need assistance in understanding their shortcomings, particularly in those cases when their grade is not necessarily low, but they are not confident in their answers. It may function as an early indicator of students not fully understanding the course material.

This grading method could be extended to increase students' actual confidence in the material, not just for assessing their answers. Engineering students tend to be relatively confident, sometimes over confident, in science and mathematics compared to non-engineering majors, who tend to be under confident (Knight and Smith 2010). For students in other disciplines taking math and science courses, an alternative grading scheme can encourage them to select confident more often, which may increase their confidence in the material. For example, awarding only 1 point for wrong and not-confident answers would increase the benefit of selecting confident.

In the long term, accurate self-assessment may be important for professional success. Knowing what one is capable of doing and learning as well as what one does not know may enable an individual to make better career decisions, ranging from selecting an appropriate project to work on to picking a specific career path. In many businesses, individuals are often promoted one level beyond where they are capable of performing well. This concept is known as the Peter Principle (Peter et al. 1969); people get promoted for doing a good job at one level, but start to struggle at the higher level because the job is distinctly different and/or they have not been trained for it (e.g., the transition from technical lead to manager). A better ongoing assessment of self and of others along with additional training efforts could reduce this problem, make employees happier, and make businesses more efficient by better matching people with the jobs in which they will be successful. The connection between these two needs further study though.

Utilizing self-assessment is context-specific since there are instances when it is appropriate to state that one is not sure of the given answer and there are cases when a lack of confidence should not be stated. The research proposed in academic proposals is inherently not guaranteed to work; to quote Albert Einstein, “If we knew what we were doing it wouldn't be research.” However, this is generally understood, so it does not need to be stated throughout the proposal. In other cases, it may not make sense to disclose that you are unsure about the proposed solution before you have more time to test it, but this relies heavily on how that solution will be used and how it was presented. One area is clear; stating one's confidence level with an answer is vital for engineering designs that affect safety and people's lives.

CONCLUSIONS AND FUTURE WORK

Engineering education typically focuses on teaching the students to use the most appropriate method and to arrive at the best answer to problems. Stating the confidence one has about an answer or design is not typically emphasized during undergraduate schooling, but it is a vital skill for jobs and in graduate education where the confidence one has in their presented work can be as important as the answers themselves. To assess student's abilities, this study integrated the CBS method into the grading of student quizzes.

Our study shows that students are reasonably good at accurately choosing their confidence selections, but could be improved. Whether or not the students learn the material better using this grading method, this method encourages students to assess their own abilities, which is a practical skill that is often overlooked in engineering education. Students generally found this method helpful and the qualitative survey indicated that students became more self-aware through this grading method. This current study did not provide any feedback nor advice about how to better assess their abilities. With this better understanding of students' baseline abilities, future studies can examine methods to improve students' abilities.

The CBS method presented here can also be used as a metric where an instructor can determine how well the students think they understand the material. As opposed to only offering additional help to students who score low, this method identifies other students who may need help in the course: those that are under or over confident. Particularly in engineering, misstating the confidence of an answer can have detrimental outcomes and, thus, self-assessment is a skill that should be encouraged.

The limitation of this and the other CBS methods described in the background is that they are only well suited for questions that are graded all-or-nothing (i.e., no partial credit grading). Although there are likely many ways to extend this method, below are three possible ways to integrate the CBS method with partial credit. The goal of these methods is to enable CBS to be implemented in problems that do not simply have a correct/wrong answer. Each of these methods tries to balance the fairness of the grading with the benefit of encouraging students to develop an understanding of their own abilities.

1. Divide each question on the test into simpler pieces with "milestones" such that students who are not sure can mark the answer to certain portions as not confident, then put down a default answer and continue on using the default answer. Each milestone would be graded as described above. This implementation does increase the effort required in grading as each problem may have two possible starting points and, thus, two answers, but it does allow each student to demonstrate the areas in which they are proficient and confident.

2. If a student marks confident, grade all-or-nothing such that the grade is 120% of its original value when correct and 0% when incorrect. If not confident, grade as normal with partial credit. This method would make grading easier since confident answers are all or nothing, but is likely to significantly increase the grades of the students that are doing very well in the course and will do little for the majority of students. A student would have to be very confident in an answer to risk this much on stating that they are confident.
3. Grade as normal with partial credit, but if not confident, the grade is scaled between 40% and 80% of the normal grade, so zero points would equate to 40% and 100% would equate to 80% with a linear transition between. This would have a minor increase in the effort required to grade, but provides a balanced benefit for many students.

Future studies that incorporate CBS should include a longitudinal study to examine if there is a correlation between subject-matter confidence and career choice. It has been demonstrated that students compare themselves to local peers, and success in their chosen field is based on their immediate peers (i.e., class/school mates) more so than peers from different institutions (Conley and Onder 2013; Gladwell 2013). This suggests that self-assessment has a large impact on students' career choices and students that could be successful in engineering may be dropping out because of a wrong perception of their abilities.

In this study, although grades do not converge throughout the semester, the ability to accurately assess one's answer did converge. Several possible studies would provide a better understanding of how and why students' assessment accuracy converges. One approach is to change the grading scheme to remove the bias toward "not confident". For example, scores for each answer/confidence combination can be changed from 5/4/2/0 to 5/4/1/0 or 5/3/2/0; these changes will likely lead to an increase in self-assessment accuracy. This change would, however, reduce the emphasis of admitting one is not confident in an answer. Another more direct approach is to require students return their graded quizzes with an explanation of two aspects for some points back: (1) if the answer is wrong, explain where and why they made an error; and (2) if they got a 0 or a 4, describe why they were overly or under confident in their answer, respectively. This method of having students respond to and return graded material is often referred to as "Exam Wrappers" (Lovett 2013). These two variations on the CBS method would help to unravel why student's self-assessment converge to a certain level over time.

ACKNOWLEDGEMENTS

Partial funding for this research was provided by the National Science Foundation under Grant No. IIS-1319802.

REFERENCES

1. Adams, J. M. (2011, December 2). "The Pessimist's Guide To Being Optimistic". Retrieved from http://www.prevention.com/mind-body/emotional-health/pessimists-guide-being-optimistic on March 3, 2015.

2. Bandura, A. (1977), "Self-efficacy: toward a unifying theory of behavioral change", Psychological review 84(2), 191. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.315.4567&rep=rep1&type=pdf

3. Bandura, A. (1997). Self-efficacy: The exercise of control. New York: Freeman.

4. Bandura, A. (2006), "Guide for constructing self-efficacy scales", Self-efficacy beliefs of adolescents 5, 307–337. http://web.stanford.edu/dept/psychology/bandura/pajares/014-BanduraGuide2006.pdf

5. Barr, D. A.; & Burke, J. R. (2013). "Using confidence-based marking in a laboratory setting: A tool for student self-assessment and learning." The Journal of chiropractic education, 27(1), 21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604960/

6. Brown, I. & Inouye, D. (1978), "Learned helplessness through modeling: The role of perceived similarity in competence", Journal of Personality and Social Psychology 36(8), 900.

7. Bruno, J. (1993). "Using testing to provide feedback to support instruction: A reexamination of the role of assessment organization," In D. Leclercq & J. Bruno (Eds.), "Item bank: Interactive testing and self-assessment," NATO ASI Series'; Berlin: Springer Verlag, pp. 190–209.

8. Carberry, A.; Lee, H. & Ohland, M. (2010), "Measuring engineering design self-efficacy", Journal of Engineering Education 99(1), 71–79. http://www.ceeo.tufts.edu/documents/journal/carberry_lee_ohland.pdf

9. Cisar, S. M.; Cisar, P.; & Pinter, R. (2009). "True/false questions analysis using computerized certainty-based marking tests," In Intelligent Systems and Informatics, 2009. SISY'09. 7th International Symposium on (pp. 171–174). IEEE.

10. Conley, J. P., & Onder, A. S. (2013). An Empirical Guide to Hiring Assistant Professors in Economics (No. 13-00009). Vanderbilt University Department of Economics. http://www.accessecon.com/pubs/VUECON/VUECON-13-00009.pdf

11. Cuddy, A.; Wilmuth, C. A.; Yap, A. J.; & Carney, D. R. (2015). "Preparatory Power Posing Affects Nonverbal Presence and Job Interview Outcomes." Journal of Applied Psychology.

12. Fantz, T.; Siller, T. & Demiranda, M. (2011), "Pre-Collegiate Factors Influencing the Self-Efficacy of Engineering Students", Journal of Engineering Education 100(3), 604–623. http://www.mychhs.colostate.edu/michael.demiranda/fantz_siller_de_miranda_jee_2011.pdf

13. Gardner-Medwin, A. R. (1995). "Confidence assessment in the teaching of basic science", Research in Learning Technology, 3(1), 80–85.

14. Gardner-Medwin, A. R., & Gahan, M. (2003). "Formative and summative confidence-based assessment," In Proceedings of the 7th International Computer-Aided Assessment Conference (pp. 147–155). http://www.ucl.ac.uk/lapt/tea/caa03.pdf

15. Gardner-Medwin, A. R. (2006). "Confidence-based marking: Towards deeper learning and better exams", in Innovative Assessment in Higher Education, Routledge.

16. Gardner-Medwin, T., & Curtin, N. (2007). "Certainty-based marking (CBM) for reflective learning and proper knowledge assessment," In REAP Int. Online Conf. on Assessment Design for Learner Responsibility.

17. Gardner-Medwin, A. R. (2008). "Certainty-Based Marking: rewarding good judgment of what is or is not reliable," Proceedings of Innovation.

18. Gardner-Medwin, A. R. (2013). "Optimisation of certainty-based assessment scores," In Proceedings of The Physiological Society. The Physiological Society.

19. Gecas, V. (1989), "The social psychology of self-efficacy", Annual review of sociology, 291–316. http://www.jstor.org/stable/2083228

20. Gladwell, M. (2013). David and Goliath: Underdogs, Misfits, and the Art of Battling Giants. New York, NY: Hachette Book Group.

21. Hassmén, P., & Hunt, D. P. (1994). "Human Self-Assessment in Multiple-Choice Testing," Journal of Educational Measurement, 31(2), 149-160. http://www.jstor.org/stable/pdf/1435174.pdf

22. Heron, G., & Lerpiniere, J. (2013). "Re-engineering the multiple choice question exam for social work," European Journal of Social Work, 16(4), 521–535.

23. Hevner, K. (1932). "A method of correcting for guessing in true-false tests and empirical evidence in support of it," The Journal of Social Psychology, 3(3), 359–362.

24. Hunt, D. P. (2003). "The concept of knowledge and how to measure it," Journal of intellectual capital, 4(1), 100–113. http://andrewvs.blogs.com/usu/files/p100.pdf

25. Jeanneret, N. (1997). Model for developing preservice primary teachers' confidence to teach music. Bulletin of the Council for Research in Music Education, 37-44. http://www.jstor.org/stable/40318837

26. Khan, K. S.; Davies, D. A.; Gupta, J. K. (2001). "Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge," Medical Teacher, 23(2), 158–163.

27. Knight, J. K. & Smith, M. K. (2010), "Different but equal? How non-majors and majors approach and learn genetics", CBE-Life Sciences Education 9(1), 34–44.

28. Kolar, H. & Carberry, A. R. (2013), Measuring Computing Self-Efficacy, in American Society for Engineering Education (ASEE) Annual Conference. http://www.asee.org/file_server/papers/attachment/file/0003/3443/5925.pdf

29. Lovett, M. C. (2013). Make exams worth more than the grade. Using Reflection and Metacognition to Improve Student Learning: Across the Disciplines, Across the Academy, 18.

30. Marra, R. M.; Rodgers, K. A.; Shen, D. & Bogue, B. (2009), "Women Engineering Students and Self-Efficacy: A Multi-Year, Multi-Institution Study of Women Engineering Student Self-Efficacy", Journal of Engineering Education 98(1), 27–38.

31. Narciss, S.; Koerndle, H.; & Dresel, M. (2011). "Self-evaluation accuracy and satisfaction with performance: Are there affective costs or benefits of positive self-evaluation bias?" International Journal of Educational Research, 50(4), 230–240.

32. Pajares, F., & Johnson, M. J. (1994). "Confidence and competence in writing: The role of self-efficacy, outcome expectancy, and apprehension," Research in the Teaching of English, 28(3), 313-331. http://www.jstor.org/stable/pdf/40171341.pdf

33. Pajares, F. (1996). "Self-efficacy beliefs in academic settings," Review of educational research, 66(4), 543–578.

34. Parsons, S.; Croft, T. & Harrison, M. (2009), "Does students' confidence in their ability in mathematics matter?", Teaching Mathematics and its Applications 28(2), 53–68. http://teamat.oxfordjournals.org/content/28/2/53.full.pdf

35. Peter, L. J.; Hull, R. & Frey, L. (1969), "The Peter Principle", Technical report, W. Morrow. http://www.msc.pef.czu.cz/msc_em/data/HarveyJ/1314/Upload3/MangementPrinciples%26Laws.doc

36. Petr, D. (2000), Measuring (and enhancing?) student confidence with confidence scores, in Frontiers in Education Conference, 2000. FIE 2000. 30th Annual, pp. T4B–1. http://erm.asee.org/pdfs/david-petr.pdf

37. Ponton, M.; Edmister, J.; Ukeiley, L. & Seiner, J. (2001), "Understanding the role of self-efficacy in engineering education", Journal of Engineering Education - Washington 90(2), 247–252.

38. Riegle-Crumb, C. & Moore, C. (2013), "Examining Gender Inequality In A High School Engineering Course", American Journal of Engineering Education (AJEE) 4(1), 55–66.

39. Rosewell, J. P. (2011). "Opening up multiple-choice: assessing with confidence," In: 2011 International Computer Assisted Assessment (CAA) Conference: Research into e-Assessment, 5/6 July 2011, Southampton, UK. http://oro.open.ac.uk/32150/1/Rosewell-poster-CAA2011.pdf

40. Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. Handbook of research on mathematics teaching and learning, 334-370. http://howtosolveit.pbworks.com/f/Schoenfeld_1992%20Learning%20to%20Think%20Mathematically.pdf

AUTHORS

Gigi Yuen-Reed is a Senior Research Scientist with the IBM T. J. Watson Research Center. She received her Ph.D. in Operations Research from Northwestern University. Her main research interests include data-driven analytics in healthcare applications, modeling healthcare ecosystem, and scalable analytics delivery platform. Prior to joining IBM Research, she was a senior manager for the Business Analytics team in IBM Global Business Services. She worked closely with financial institutions and healthcare companies to enable analytics technologies as an integral part of business solutions. She designed and led various engagements that bring the latest innovations from Research and Software to market. She also serves as an adjunct professor with the Department of Industrial and Management Systems Engineering at University of South Florida.

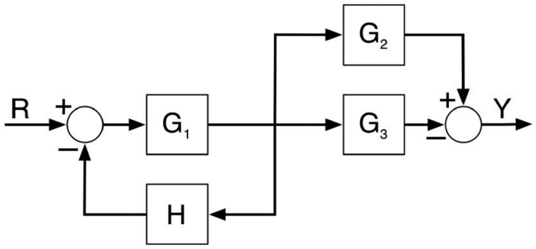
Kyle B. Reed is an Assistant Professor of Mechanical Engineering at the University of South Florida. He received the B.S. degree in Mechanical Engineering from the University of Tennessee in 2001 and the M.S. and Ph.D. degrees in Mechanical Engineering from Northwestern University in 2004 and 2007, respectively. He was a postdoctoral fellow in the Laboratory for Computational Sensing and Robotics at The Johns Hopkins University. His research interests include haptics, human-machine interaction, rehabilitation engineering, medical robotics, and engineering education.

APPENDIX: EXAMPLES OF SPECIFIC QUIZ QUESTIONS

The following are examples of the quiz questions asked on the quizzes used in this study.

Question 1 as listed in Table 2 (from Quiz 2):

Find $\frac{Y}{R}$ (5 pts.): ____ not confident



Question 3 as listed in Table 2 (from Quiz 3):

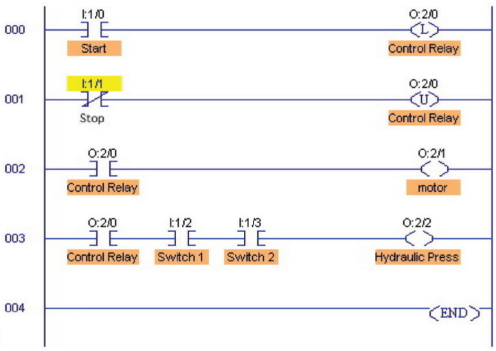
Use the final value theorem to find the value of $f(t)$ at $t = \infty$ for $F(s) = \frac{7s + 27}{s^2 + 6s + 9} * U(s)$ where $U(s) = \frac{1}{s}$ is the unit step input.

____ not confident

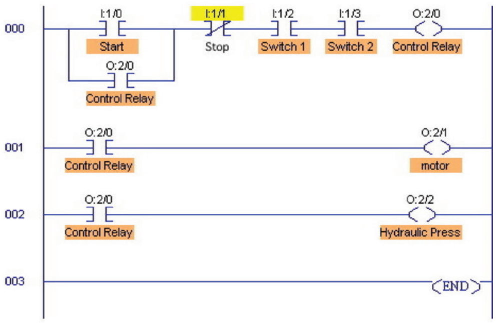
Question 14 as listed in Table 2 (from Quiz 6):

Consider a hydraulic press system which has a motor to pump fluid and two switches to activate the hydraulic press. The *motor* pump should start when the *Start* button is pressed and turn off when the *Stop* button is pressed, even if the *Hydraulic Press* is not on. For safety reasons, two separate push-button switches (*Switch 1* and *Switch 2*) need to be closed simultaneously to activate the *Hydraulic Press*. Circle the one ladder diagram below that would best perform as stated above?

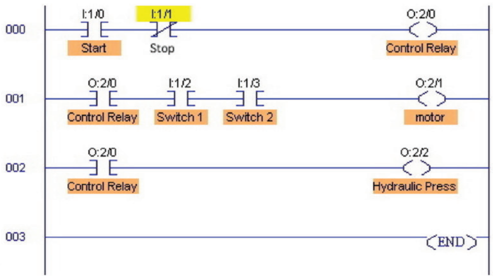
not confident ____



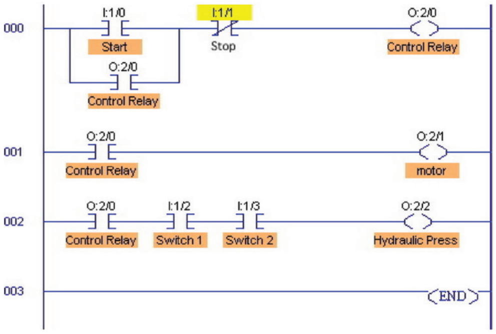
a)



b)



c)



d)