# MEASURING (AND ENHANCING?) STUDENT CONFIDENCE WITH CONFIDENCE SCORES

*David W. Petr*[1]

***Abstract** -- The important skill of building confidence in one's analysis through sanity- and cross-checking is often poorly acquired by engineering students. An introductory circuit analysis class presents an ideal opportunity in which to emphasize and measure this skill, since problems can typically be worked with a number of different methods or worked "backwards" to provide cross-checks. This paper reports on a two-semester experiment in which students were required to provide a confidence rating for all exam and quiz answers. The structure allowed for expressing positive confidence (confidence that an answer is correct), negative confidence (confidence that an answer is incorrect), and neutral confidence. A confidence score that measured how well the students evaluated the correctness or incorrectness of their answers was combined with a traditional problem score to form the exam score. We present numerical results of this experiment, which yield potentially valuable conclusions regarding students' perceptions of the correctness of their answers.*

*Index Terms – assessment, confidence scores, results-checking, student confidence*

## INTRODUCTION

The important skill of building justified confidence in one's analysis through sanity- and cross-checking is often poorly acquired by engineering students. Many students are confident in their analytical results, but it is often a false confidence that fails to recognize the very real possibility of a host of errors including improper application of theory, incorrect problem formulation, inaccurate data entry, and improper use of calculators and computers. Indeed, the problem has been exacerbated by our reliance on modern computational tools, whose "infallibility" is often tacitly extended to the user of such tools. As observed by a mathematics lecturer interviewed on this subject [1], "They see the calculator as something that will solve the problem for them, rather than as an aid to solving the problem. This is a major problem. Because a number has been generated on a calculator, they are convinced that it must be right." To become successful engineers, students need to develop a healthy skepticism toward numerical results that prompts them to ask questions such as "Does this result make sense?" (sanity-checking) and "How can I confirm it?" (cross-checking) before proceeding further.

Although some sanity-checking requires experience with or a thorough understanding of a subject, an introductory circuit analysis course presents an ideal opportunity to emphasize simple sanity-checking and especially cross-checking early in an engineering curriculum. Circuit analysis is rich in conservation laws that can be applied to results and in alternative methods that can be used for re-working problems. Human nature being what it is, however, students need some incentive for applying these techniques.

This paper describes a two-semester experiment in which confidence scores were used on exams and quizzes to encourage students to use sanity- and cross-checking and to measure their confidence levels. The following sections describe experimental background, the methodology used, numerical results, other considerations, conclusions, and recommendations.

## EXPERIMENTAL BACKGROUND

This experiment was conducted in two consecutive semesters (Spring and Fall of 1998) of a five-hour introductory circuit analysis class at the University of Kansas. The course was intended primarily for second-year electrical or computer engineering students, although significant numbers of aerospace engineering students of various levels also enrolled. A total of 80 students participated, 44 in the first semester and 36 in the second.

Four exams were given in each semester. Corresponding exams in the two semesters covered very much the same material. Since the confidence scoring (described below) was expected to require additional time beyond working out the problem solutions, extra time was given for the exams. The first three exams were constructed as though they would be taken in an 80-minute lecture period, but the students were allowed 120 minutes (a separate evening meeting) to complete each exam. The final exam was constructed for a 2-hour period, whereas students were actually allowed 3 hours. In addition, seven short (15-minute) in-class quizzes were given in the first semester, of which the last six included confidence scores. Again, an attempt was made to allow ample time for completing the quiz problems and the confidence scoring.

[1] David W. Petr, University of Kansas, Department of Electrical Engineering and Computer Science, Lawrence, KS 66045, petr@eecs.ukans.edu

## METHODOLOGY

In addition to emphasizing the use of sanity- and cross-checking in the lectures and homework assignments, a means was required to encourage the students to put these techniques into practice on exams and quizzes. The methodology should encourage the development of "physical insight and judgement" (sanity checking) and "the inclination and ability to check one's work" (cross-checking) [2]. Several means were considered. One could somehow require written justification of confidence such as evidence of cross-checking. One could introduce a few problems specifically aimed at confidence checking, such as giving a problem and an answer and asking if the answer is correct. Another approach [2] uses a subset of short, carefully designed questions that are graded "all or nothing" – either full credit or no credit – thereby providing extra motivation for the student to ensure that the answer is indeed correct.

In this research, we desired a means by which students would be rewarded for taking a relatively small amount of time and effort to properly evaluate their own numerical answers (and penalized for not doing so). Certainly students should be rewarded for recognizing a correct answer. Just as importantly, however, they should be rewarded for recognizing an incorrect answer, since this is the first step in correcting the error. As the students were told, "You are much less dangerous if you are wrong and know it than if you are wrong and oblivious!" We also desired a means to explicitly measure their confidence levels and to track trends in their confidence levels with time. Finally, we desired a methodology that would require neither an excessive amount of the students' time to complete nor an excessive amount of the instructor's time to grade.

To these ends, the following confidence scoring system was developed. Each exam and quiz consisted of two parts, a problem part and a confidence part. The problem part was identical to this instructor's typical exam or quiz. Problem statements were given (with a point weighting for each distinct answer), and the students were instructed to work the problems in the space provided, showing steps and clearly indicating their final answers. The problem part was graded in this instructor's typical fashion, granting partial credit for the steps shown even if the final answer was not completely correct (including units). The result was a total problem score for each exam.

The confidence part of the exam or quiz was the new part and the focus of this paper. A separate answer/confidence sheet (see Figure 1) was provided for the confidence part. Students were instructed to copy each distinct answer (including units) to the corresponding space on the answer/confidence sheet. In addition, students were instructed to "Indicate your confidence rating for each answer by circling a number from –5 (very confident it is wrong) to 5 (very confident it is correct)." Negative confidence ratings thus represented a sort of negative confidence, and a neutral confidence rating of 0 was to indicate that the student had no idea about the correctness of the answer.

| Answer | Confidence (+5 is high) |
|---|---|
| 1a. | -5 -4 -3 -2 -1 0 1 2 3 4 5 |
| 1b. | -5 -4 -3 -2 -1 0 1 2 3 4 5 |

FIGURE 1
FORM OF ANSWER/CONFIDENCE SHEET

For each answer, a confidence score ($C$) was computed as

$$C = V \cdot W \cdot (R/5)$$

where $V$ is +1 if the answer given on the answer sheet was completely correct (including units) and –1 if not, $W$ is the answer's point weighting, and $R$ is the answer's confidence rating. In this way, the confidence score was positive if the student believed the answer to be correct and it was or if the student believed the answer to be incorrect and it was. A mismatch between the student's view of the correctness of the answer and the actual correctness of the answer resulted in a negative confidence score. The confidence score could thus range from $W$ (the answer's weight) to $-W$. All weights were integer multiples of 5 to ensure that all confidence scores were integers, making it feasible for this grader to compute confidence scores mentally. The total confidence score was the sum of the individual confidence scores.

The preliminary exam score ($E_P$) was computed as

$$E_P = 0.98 \cdot P_T + 0.05 \cdot C_T$$

where $P_T$ is the total problem score and $C_T$ is the total confidence score. For all results presented in this paper, the scores $P_T$ and $C_T$ are normalized to have a maximum value of 100, and $E_P$ was considered as a 100 point scale. This form for $E_P$ allows students to be rewarded with at most three points "extra credit" (on a 100-point basis) for perfect confidence scores. In addition, they could be penalized up to two points for refusing to give confidence scores or making all the confidence ratings 0, and penalized at most 7 points for completely inaccurate confidence ratings (since $C_T$ has a minimum value of –100).

The final exam score ($E$) was then computed as

$$E = E_P + A$$

where $A$ is a constant grading adjustment (for a given exam) applied to every student. Exam grades as reported here were determined by applying the common 90%, 80%, 70%, 60% thresholds to the exam grade $E$.

## NUMERICAL RESULTS

### Confidence Scores by Student Performance

From the construction of the confidence scores, one might expect that very good students and very poor students would have the highest confidence scores, since they would be more likely to accurately judge the correctness or incorrectness of their answers. Figure 2, which plots average confidence score (on a 100-point basis) as a function of student exam or quiz grades, partially supports this hypothesis. One interesting feature of Figure 2 is the consistency in trends, especially between the exam curves for the two semesters. While the "A" and "F" students are indeed the two highest scoring groups in each case, note that the "A" students significantly outscore the "F" students, who score only marginally better than the other non-"A" students.

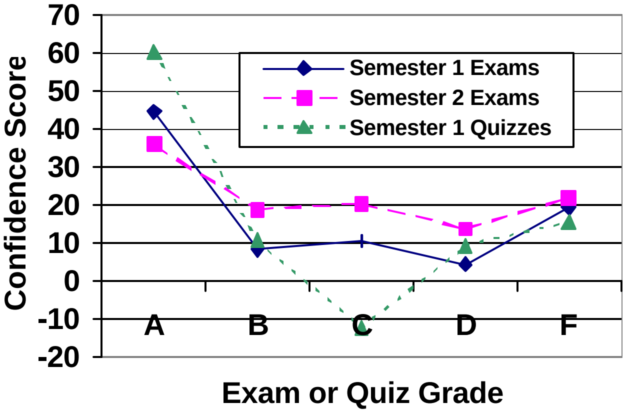

FIGURE 2
CONFIDENCE SCORES VS. PERFORMANCE

A related topic, and one of particular interest to students, is whether this confidence scoring method favors some students more than others. Figure 3 plots the confidence score bonus, defined as the difference between the exam score $E_P$ calculated using the confidence scores and the conventional exam score $P_T$, as a function of exam or quiz grade. Because of the way $E_P$ is calculated, the higher-scoring students have to get higher confidence scores in order to get the same bonus as lower-scoring students, so Figure 3 is a skewed version of Figure 2. Figure 3 shows that the use of confidence scores slightly increased the exam/quiz scores of "A" students (by about 1 point or less on a 100-point scale) and slightly decreased the exam/quiz scores of "B" and "D" students (by about the same amount). Confidence scoring had a small negative effect on the exam scores of "C" students and a somewhat larger negative effect on the quiz scores of "C" students. Confidence scoring had little statistical effect on the exam/quiz scores of "F" students.

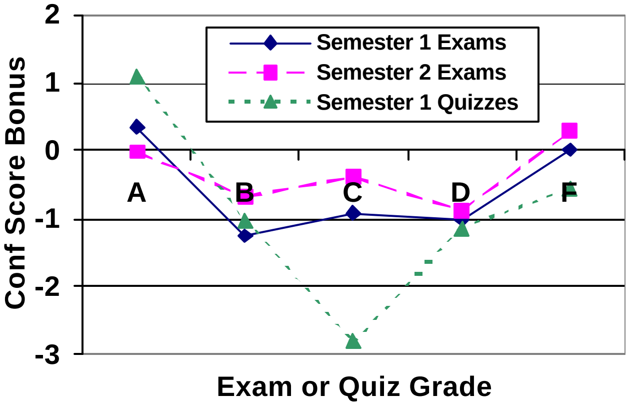

FIGURE 3
CONFIDENCE SCORE BONUS VS. PERFORMANCE

### Confidence Scoring Over Time

An important question is whether or not students improved in their ability to properly assess their answers after repeated practice with confidence scoring. Student confidence scoring as a function of time, as shown in Figure 4 and Figure 5, is one way to measure this. Although neither figure is particularly conclusive, there does seem to be a general upward trend after the first exam or quiz. The especially high confidence scoring on the first exam each semester may be related to the students' especially good overall scores on this exam, which were 9 and 5 points better (out of 100, respectively) than any of the later exam scores. Also, the material covered in the first exam (simple resistive circuits including Ohm's Law, power conservation, Kirchoff's Laws, simple nodal and mesh analysis) is quite rich in sanity- and cross-checking options. Note again the similarity between the two semesters in Figure 4.

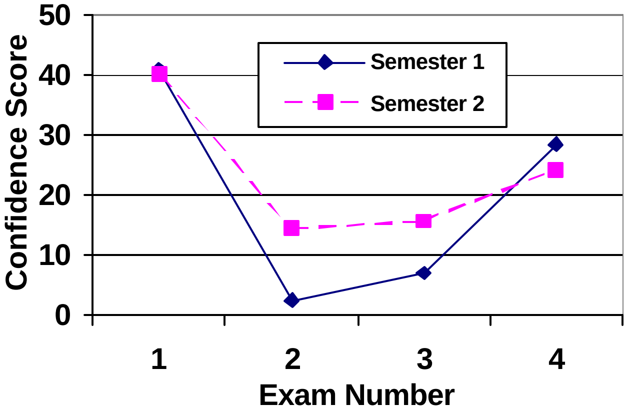

FIGURE 4
EXAM CONFIDENCE SCORES OVER TIME

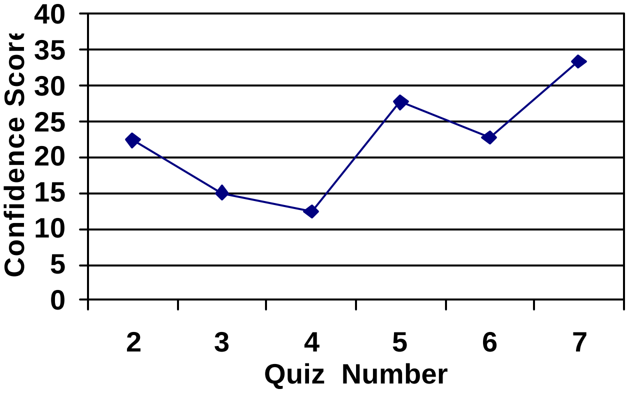

FIGURE 5
QUIZ CONFIDENCE SCORES OVER TIME

### Confidence Matching Categories

The generally poor confidence scoring by the non-“A” students (Figure 2) leads one to wonder if students are more confident in their abilities than they should be, thinking that their answers are correct when they are not. This was investigated using answer-by-answer data gathered from each semester’s final exam. In Figure 6, we show the distribution of the student confidence ratings on the final exams among five confidence matching categories: True Positive (TP), False Positive (FP), Neutral (N), False Negative (FN), and True Negative (TN). In this categorization, positive/negative indicates the sign of the confidence rating and true/false indicates whether the student’s view of the correctness of the answer did or did not match the actual correctness of the answer. Neutral indicates a confidence rating of 0. Several important points emerge from Figure 6.

First, a relatively large percentage of the answers carried a neutral confidence rating (29% in semester 1 and 21% in semester 2), indicating that the student had insufficient motivation or time to carry out any sort of sanity- or cross-checking for that answer. Given that this data is from the final exam in the course, this is a discouraging result. However, it should be noted that students frequently have a good idea of the minimum score required on the final exam to obtain a given overall grade in the course, thereby reducing their motivation to score as high as possible on the final exam.

Figure 6 also shows that positive confidence is more frequently chosen than negative confidence but is also much more frequently false than negative confidence. That is, it is much more common for a student to think an answer is correct when it is not (false positive confidence) than it is for a student to think an answer is wrong when it is correct (false negative confidence).

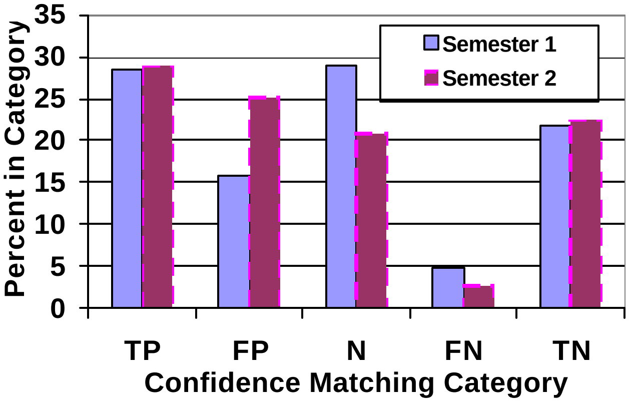

FIGURE 6
DISTRIBUTION OF CONFIDENCE MATCHING CATEGORIES

Figure 7 provides more information about confidence matching categories by showing the mean confidence rating (magnitude) as a function of confidence matching category. Here we see a strong tendency to choose extreme confidence ratings (+5 or –5 was selected 46% of the time in semester 1 and 67% of the time in semester 2), indicating that students were not willing or able to make fine distinctions in their levels of confidence. In fact, students chose +5, -5 or 0 for their confidence rating 75% and 87% of the time in the two semesters. It is also curious to note the consistent tendency of students in the second semester to choose more extreme confidence ratings. Further, although we might expect to see significantly smaller confidence ratings when the confidence is false, Figure 7 shows that this is only marginally the case.

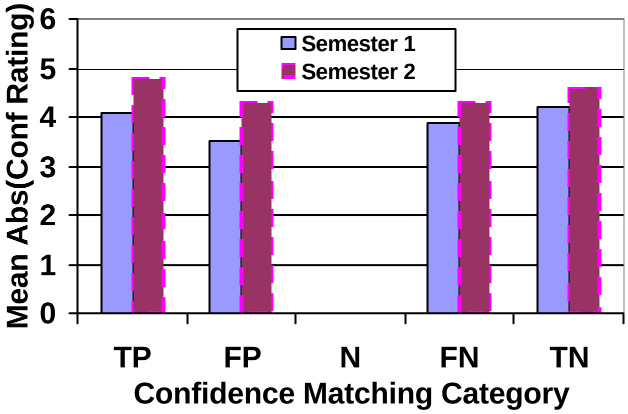

FIGURE 7
MEAN CONFIDENCE RATINGS BY CATEGORY

### A Possible Modification

It could be argued that the means of computing the confidence score $C$ for each answer is biased against the students because of the “all or nothing” nature of the validity variable $V$: $V$ is +1 only if the answer is entirely correct and –1 otherwise. The generally poor confidence scores of the non-"A" students (Figure 2) may be evidence of such bias. The binary definition for *V* was chosen in part to keep the computation required of the grader at a manageable level, but alternative forms for *V* could be considered.

One approach would be to replace the binary nature of *V* with a calculation that reflects the degree of correctness of the answer. In this way, a student who gives an answer that is nearly (but not completely) correct, perhaps due to an arithmetic or units error, would not be penalized as harshly for believing that the answer is indeed correct. One possibility would be to calculate *V* as

$$V = \frac{2 \cdot P - W}{W}$$

where *P* is the answer's problem score and *W* is the problem weight, as before. This definition allows *V* to take on a range of values between +1 and –1, expressing the degree of validity (or correctness) of the answer.

Unfortunately, there are some serious problems with this approach. First, it is more computationally burdensome for the grader, probably requiring entry of each individual answer into a spreadsheet. Second, this method may encourage a more subjective approach to rating confidence, rather than a more objective approach based on sanity- and cross-checking, which tends to produce binary results. Finally, using the existing database of confidence ratings and problem scores for the final exams, Figure 8 (compare to Figure 2) shows that this alternative method of computing the confidence score has the effect of making the confidence scores more closely correlated with problem scores, thereby favoring the better students. However, there is a danger in applying this alternative definition to the data at hand, since the students were making their confidence decisions based on the original, binary definition of *V*. A change in the definition of *V* may well change the students' choices of confidence ratings.

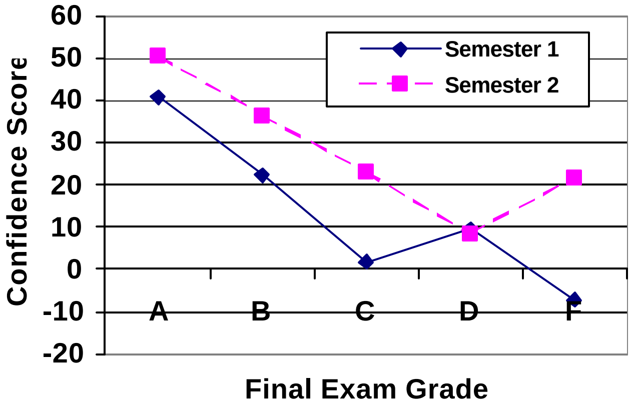

FIGURE 8
MODIFIED CONFIDENCE SCORES VS. PERFORMANCE

## OTHER CONSIDERATIONS

### Student Reaction

Students were specifically encouraged to comment on the confidence scoring idea and procedures in their end-of-semester evaluations of the course, but only eight of the 80 did so (56 of the 80 provided some written comments). This relative silence is difficult to interpret, but may indicate apathy or ambiguity towards the concept.

Of the eight who commented, six expressed dislike for the confidence scoring, with reasons such as "wasn't really relevant," "dragged down my grades," "ineffective…and un-useful in learning material." Of the six expressing dislike, one tempered his/her dislike with "it wasn't worth much on tests and you did support yourself well on why you did it" and another commented: "Yes, we do hate [the] confidence score, but I think it's a good idea. Keep it." Only one student said "I like the idea," perhaps in part because "[I] benefited on 2 out of 3 tests because of it."

Two students expressed frustration with the requirement of copying their answers to a separate page, both mentioning the possibility of transcription errors. One student suggested a coarser confidence scale of -2, -1, 0, 1, 2. One mentioned the possibility of losing "the whole confidence score" because of math or transposition errors, a comment that partially motivated the investigation above of a modified confidence score.

### Grading Time

No attempt was made to measure the extra time required to process the answer/confidence sheets and compute the exam score from problem and confidence scores. The extra time could be characterized as significant but not substantial; a very rough estimate would be 10% of the time that was required to determine the problem score. Extra grading time would be reduced slightly with a coarser confidence scale.

### Some Difficulties with Cross-Checking

One might question (the students certainly did!) whether there was really sufficient time allowed within the limited examination times to do any significant cross-checking, in particular working a problem with a different method. Also, if a student works a problem with two different methods and obtains the same answer from each, certainly that should be cause for positive confidence in the answer. But what should the student's response be if two different answers are obtained? Obviously they can't both be correct, but one might be. In such a case, the confidence rating probably reduces to a subjective judgement concerning the student's relative confidence in the particular methods involved, for example being more "comfortable" with mesh analysis than with nodal analysis.

## CONCLUSIONS

We can draw some conclusions about student confidence and about this particular confidence scoring methodology, based on the results of this experiment.

- "A" students were significantly better at evaluating the correctness or incorrectness of their answers than were any other grade-based class of students.
- Relative to traditionally scored exams, this confidence scoring methodology tended to favor slightly the "A" students, slightly hurt the "B", "C", and "D" students, and be neutral towards the "F" students.
- There is weak evidence that students became better at evaluating the correctness or incorrectness of their answers as the semester progressed. There is no way of determining from this experiment whether the use of confidence scoring had a positive or negative effect on students' confidence-building skills.
- Even after a semester of emphasis, a significant number of the answers (approximately 25%) were assigned a neutral confidence rating, indicating an unwillingness or inability of the student to determine that the answer was either correct or incorrect.
- Students were nearly twice as likely to believe their answers were correct (positive confidence) than they were to believe that their answers were incorrect (negative confidence).
- Students tended to be optimistic about the correctness of their answers. That is, answers believed by students to be correct often were not (about 40% of the time).
- Answers believed by students to be incorrect usually were (about 85% of the time).
- Students were unable or unwilling to make fine distinctions in confidence ratings, choosing extreme or neutral ratings approximately 80% of the time.
- Students were mildly antagonistic toward the confidence scoring system.

## RECOMMENDATIONS

This experiment has yielded potentially valuable information about student confidence, but it is not recommended that the confidence scoring methodology presented here be adopted for use on a routine basis. Student animosity, extra student time required for the procedure, and extra instructor time for grading and record-keeping all argue against such a use. Instead, other procedures, such as a combination of short questions encouraging sanity- and cross-checking combined with longer problems [2], accomplish some of the same purposes with fewer drawbacks.

However, the confidence scoring methodology could be used in specific experiments such as this for gathering further data about student confidence levels. For any such future experiments, here are some recommendations.

- Continue to use confidence scoring in a low-level circuit analysis or similar course.
- Reduce the confidence rating possibilities to three: -1, 0, and +1.
- Allow considerable extra exam time when using confidence scoring, perhaps as much as twice the time normally allotted.
- Secure assistance with the grading and record-keeping. The answer/confidence sheets can be easily graded by someone with little or no knowledge of the subject matter.
- For data analysis purposes, keep detailed, problem-by-problem records of problem and confidence scores for all exams, similar to what was done on the final exams in this experiment.

## REFERENCES

1. Tizard, Jenny, "Will Core Skills Improve Engineering Programs?" *International Journal of Electrical Engineering Education*, Vol. 32, No. 2, 1995, pp. 99-107.
2. Hanrahan, H.E., "Effective Examining Using Short Questions," *International Journal of Electrical Engineering Education*, Vol. 29, No. 3, 1992, pp. 205-211.