2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #53695

WIP: Self-reported certainty in engineering students' knowledge of bioinstrumentation and validity of statements from peers and LLMs

Joseph Tibbs, University of Illinois at Urbana - Champaign

Joseph Tibbs is a Bioengineering PhD candidate at the University of Illinois Urbana-Champaign. He is interested in pursuing a passion for education by becoming a teaching faculty and exploring topics of ethics education, engineer identity formation, and bioengineering curriculum development.

Kaitlyn Tuvilleja

Kaitlyn Tuvillea is a Bioengineering Undergraduate with a Statistics Minor at the University of Illinois at Urbana-Champaign. Her interests in the therapeutics, cell and tissue engineering, and bioengineering curriculum development are complemented by her involvement in Society of Women Engineers (SWE), Women in Engineering (WIE), Biomedical Engineering Society (BMES), and Brain Matters.

Dr. Rebecca Marie Reck, University of Illinois Urbana-Champaign

Rebecca M. Reck is a Teaching Associate Professor of Bioengineering at the University of Illinois Urbana-Champaign. Her research includes alternative grading, entrepreneurial mindset, instructional laboratories, and equity-focused teaching. She teaches biomedical instrumentation, signal processing, and control systems. She earned a Ph.D. in Systems Engineering from the University of Illinois Urbana-Champaign, an M.S. in Electrical Engineering from Iowa State University, and a B.S. in Electrical Engineering from Rose-Hulman Institute of Technology.

WIP: Self-reported certainty in engineering students' knowledge of bioinstrumentation and validity of statements from peers and LLMs

Abstract

Laboratory courses seek to translate classroom knowledge to applications in the real world. In that translation, students must grapple with complex problems in which they learn that there is more to learning than simple factual recall. The ability to gauge their own confidence in knowledge, or the ability to admit when they don't know something, is essential to their future career when their work will be relied upon to improve or save lives. When students are asked to evaluate their knowledge, and whether they are certain about it, they explore their thought processes critically and deepen their learning. We have previously done work on students' confidence in their ability to learn relevant material for a bioinstrumentation laboratory course, and we have improved the feedback methods to better provide scaffolded, formative assessment. This work builds on both of these concepts to develop a model of student certainty, or their self-evaluation of performance on a specific knowledge task, while also studying student epistemologies in engineering. We are implementing this in a laboratory course setting because there is often a gap between pre-lab assignment knowledge and the critical thinking necessary to use that knowledge in an instrumentation context.

The purpose of this study is to develop a formative assessment activity which introduces students to the concepts of certainty and accountability in engineering practice, while also confronting them with statements of varying subjectivity to prepare them for poorly-defined problems. Students are given short ungraded quizzes at the beginning of their lab section which ask them to respond with an answer to a question, an assessment of their certainty in that answer, and an evaluation of the validity of a statement derived either from a peer or an online source. This is aligned with the course objective of promoting critical thinking about course content, but will also provide quantitative data on how reported certainty is affected by the source of the statements they evaluate. By asking the same question in three different ways, we can determine if students react differently to information that comes from their own memory, a peer, or a Large Language Model. We will determine whether student certainty correlates with characteristics like declared major or reported gender, and whether certainty correlates with correctness.

Introduction

Large Language Models (LLMs) are tools which have become broadly available in recent years, and their prevalence has dramatically changed the landscape of education and assessment [1] [2]. There are immediate concerns about the usage of these tools for academic dishonesty, as a more accessible means to create plagiarized written work than traditional paid essay services. However, these tools exist on a spectrum of use cases for text processing and manipulation; students may use them to brainstorm or draft outlines of work, proofread or polish a finished paper to improve its clarity, or engage with an LLM as an interlocutor for effective one-on-one tutoring [3]. Thus, professors are rarely banning all interactions with LLMs, only those that significantly impact the students' ownership over content that is produced or knowledge that is referenced. This is because LLMs are *language* models, not *knowledge* models, resulting in the

phenomenon of "hallucination" or statements that are grammatically correct while being factually misleading. It therefore becomes important to evaluate what LLMs know, or what they can claim to know. However, the practical implication of this quality is not *whether* an LLM hallucinates but how a human operator reacts when hallucinations happen [4]. What can we learn about human knowledge, and students' learning, based on their interactions with a source of convincingly worded but unreliable information?

Some researchers are already examining how these interactions might play out in a classroom setting. A student with access to the free public version of ChatGPT may entirely rely on the LLM's ability to solve homework problems in a standard undergraduate engineering course with minimal editing and receive a passing grade [5]. This calls into question the current design strategies for assessment of knowledge, both to improve fairness for grading to reward students who are truly learning and to acknowledge that the goals of education may be shifting now that the landscape of tools available for problem-solving has expanded. The advent of pocket calculators de-emphasized the need for rote computation as an education outcome; what skills should educators emphasize now that undergraduate-level quantitative analysis can seemingly be automated?

This study seeks to explore one possible avenue for knowledge-based skills that future engineers must develop in their training: certainty [6]. This is different from confidence, which is an affective judgment of one's ability to perform abstract tasks. Certainty, instead, is the assessment of whether a student believes their answer is justified when given a specific objective question. Importantly, it is also the ability to honestly acknowledge ignorance when they cannot sufficiently justify the answer. Practically speaking, it is the difference between being able to make decisions that hinge on the answer or needing to seek a second perspective on the problem before moving forward. "Knowledge" is the lowest tier on Bloom's Taxonomy of learning tasks; being able to articulate the informed absence of knowledge is more difficult [7]. It requires critical thinking and a deeper analysis of what their knowledge is based on. Put simply, do students know what they don't know? And would the development of that skill help to protect them against the hallucinations of now-ubiquitous "AI" output? And is there an inherent benefit to gauging certainty of knowledge along with its retention?

This is not a new question. In 1936, Soderquist reported a simple method of weighted scoring based on students reporting the certainty with which they answered each question: "Thus the examination was calculated to test not only knowledge but also the certainty with which the knowledge was held" [8]. Other authors from the early 20th century similarly emphasized the importance of students being able to accurately express certainty in what they know, especially for fields such as medicine and engineering where important decisions will be made based on that knowledge [9]. By assessing certainty, multiple-choice test results could also be made more reliable and less dependent on guessing. However, there were valid criticisms of these test methods: students found them complex, unwieldy, or disconcerting, especially in cases where confident wrong answers were heavily penalized. Researchers were concerned that differences in student attributes (such as self-esteem or loss-aversion) would impact the ability of scoring systems to operate equivalently for all students, especially if there were a gendered component to these differences [10]. Indeed, some quantitative studies used survey instruments designed to measure personality components such as "risk-aversion" [11] in conjunction with certainty

estimates on an objective assessment, and found correlations in this variable with the tendency to over- or under-estimate certainty in a response [12]. However, other studies found that attributes of this type were significant only when students were relatively unfamiliar with the certainty-based component of scoring [13] [14], that these differences disappeared for sufficiently large sample sizes of routinely-applied certainty metrics [15] [16], or that individual characteristics were primarily relevant when the incentive structure for inaccurate answers carried a significant score penalty and thus motivated students to respond "strategically" [17]. Gardner-Medwin has reported extensively on the mathematical theory underpinning a large dataset of certainty-weighted true-false questions administered consistently over decades in a medical curriculum at the University College London [18] [19]. His conclusion is that the critical thinking encouraged by self-reflection during the answer process is valuable to learners and informative to educators. Certainty assessment metrics have also been used in laboratory settings [20] and engineering classrooms [21] and are an active area of study for summative assessment in high school settings to instill metacognitive habits in students [22].

In contrast to those examples from the literature, the focus of this study is on formative assessment. The emphasis on feedback mechanisms in this class is designed to help students be more reflective and learn more deeply, rather than focusing on a grade as a measurement of achievement [23]. This is aligned with the advice to replace high-stakes one-time assessment with more continuous evaluation of learning in the modern age of AI-aware learning [24]. By introducing a step where students evaluate their learning, every assessment becomes an opportunity to strengthen their mastery of the material by helping them realize what they still need to learn. As shown in the results, the data is also a useful tool for educators who wish to design effective questions to promote student understanding. By giving students the opportunity to identify what they don't know, honestly acknowledged ignorance can become a normalized part of the learning environment, while students develop genuine security in the knowledge they construct from a variety of sources.

Framework

Figure 1 shows the conceptual framework being used to evaluate the data collected in this study. The theoretical basis of the analysis performed here is founded, first, on the idea that knowledge is internalized by students in a constructivist manner [25]. When they are asked to give an answer to a problem, they are generating a new statement (the question combined with its answer); their certainty in the correctness of the response is the result of comparing it to an internal reference model to see how well it connects to existing knowledge. If they can generate many connections to support the statement, they will have high certainty in the answer. However, as previously stated, the decision of how to report that certainty can be impacted by individual characteristics such as risk attitude or test anxiety in addition to knowledge mastery [13] [17]. We therefore assume that some modification of constructed certainty occurs based on statement context, social consciousness, assumed expectations, personality, and other internal variables that cannot be directly measured. The result is a certainty statement that the student is willing to report in a self-evaluation question on an assessment, emphasizing their reflection.

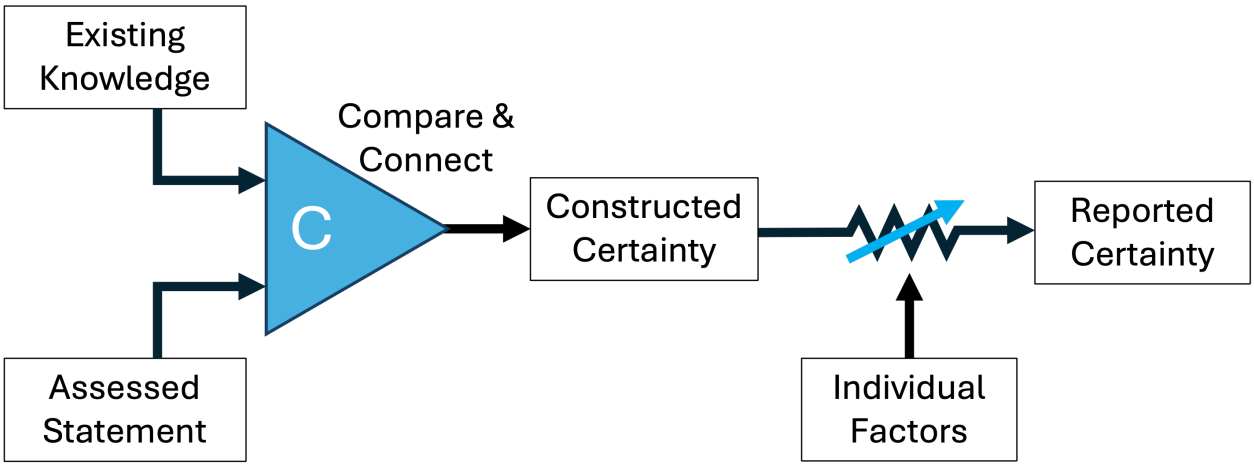

Figure 1: **Conceptual framework for study analysis.** Student certainty is an internal variable based on a comparison of a statement or question response against their existing model of knowledge and any connections they can make which confirm or disconfirm that statement. The certainty that students report is then modified from that starting point based on factors dependent on the question context and student characteristics.

Methods

Course context: The study was conducted at a large public midwestern university. Students in the study were enrolled in the BIOE 415/ECE 415 Bioinstrumentation Laboratory course. Most students were in their third year, and most were Bioengineering majors who were taking the course as a requirement, with some Electrical/Computer Engineering students choosing the course as an elective. Students attended a one-hour lecture to introduce concepts relevant to the lab, but much of student learning occurred prior to classroom interaction as they reviewed textbook materials to complete the pre-lab quiz before lecture. In the laboratory portion, students completed hands-on pre-written lab assignments and wrote reports on their results. For most of the BIOE students, an accompanying lecture course (BIOE 414) was required as a more thorough grounding in electronics and circuit concepts. In the first semester of data collection, N=31 students consented to have their quiz responses analyzed and used for this research. The following semester, for which data collection is ongoing, has significantly more students enrolled.

Each assessment took the form of two-question "minute-quizzes" administered at the beginning of class as a transition from the announcement period to the independent work section. Students were told that these quizzes would not affect their grade. Quizzes were physically printed on half-sheets of paper and handed to students, to be turned in as soon as they were complete. Quiz questions were designed to be aligned with course objectives (each lab assignment has a list of course objectives or essential skills associated with it) and related in some way to the material covered by the pre-lab. Potential quiz questions were administered to members of the research team who had previously taken the course to gauge whether the questions seemed fair and properly aligned.

A)

1. True or false: “Because the power source has a floating ground, the voltages that the power source outputs will be relative to the COM port, not necessarily the GROUND port”

Answer:	I am: a. Not certain at all; this is a guess b. Somewhat certain c. Mostly certain d. Completely certain; I would sign my name to this answer

B)

2. You ask ChatGPT for help on understanding your lab 1 report and it outputs the following sentence: “Because the power source has a floating ground, the voltages that the power source outputs will be relative to the COM port, not necessarily the GROUND port”
   a. This is definitely wrong
   b. I think this is wrong or misleading
   c. I don’t know whether it is right or wrong
   d. I think this is correct but I’m not sure
   e. I know this is correct and I would sign off on it

C)

2. You are writing the lab 1 report with a classmate and you see that they have written the following sentence in your shared document: “Because the power source has a floating ground, the voltages that the power source outputs will be relative to the COM port, not necessarily the GROUND port”
   a. This is definitely wrong
   b. I think this is wrong or misleading
   c. I don’t know whether it is right or wrong
   d. I think this is correct but I’m not sure
   e. I know this is correct and I would sign off on it

Figure 2: **Quiz questions illustrating parallel question design.** Quiz questions from three versions of the quiz given to students at the sixth lab session. The True-False question from the knowledge check section (A) was repurposed as a validation task for sources of ChatGPT output (B) and peer work (C). Responses and certainty for all question types were then analyzed to examine any systematic differences in student self-reflection.

For each week's assessment, four versions of the quiz were designed with different questions; these were distributed in a pseudo-random fashion such that students sitting next to each other did not receive the same version of the quiz (to discourage collaboration). Although the questions were different, each quiz had the same format: the first question (which may be true-false, multiple choice, or free-response) asked for an answer (knowledge check) as well as a certainty rating on a scale of four options (see Figure 2A). The second question was presented as a statement that students would evaluate as being true or false (statement validity) on a scale of five certainty options (see Figure 2B and 2C). For the second question, there were two designs: statements that were objectively true and statements that were incomplete or ambiguous. As can be seen in Figure 2, the objective statements were taken directly from a true-false question 1 on a different version of the quiz, and were simply presented as being output from ChatGPT (Figure 2B) or a contribution from a peer (Figure 2C). The four versions of each quiz meant that students would encounter the statement only once, in one of these three contexts.

To analyze the data, student names were replaced by unique numerical identifiers, associated only with self-reported characteristics such as major and gender. For the purposes of gender categorization (to prevent groups with a small number of members from being either excluded or singled out in a way that would violate the privacy of their data), a dichotomous structure of student gender as either "dominant gender in engineering" or "non-dominant gender in engineering" was used, with students identifying as male belonging to the first group and all others to the second.

To evaluate the between-group differences in responses for different question types, the range of responses were coded to a numerical spectrum and compared using the Kolmogorov-Smirnov test. For instance, the set of certainty ratings claimed by students who got a question wrong might be compared to the ratings from students who answered the same question correctly. If certainty has a bearing on question performance, a K-S test would be able to differentiate the two distributions. However, student certainty ratings are also subjective and may be affected by the wording of the categorical descriptors and personal characteristics. To normalize for some of these effects, an alternative post-hoc certainty scale was determined by converting each students' certainty to a personal "z-score" of their certainty: c_z. This was based on the mean and standard deviation of each individual student's certainty when responding to the knowledge check questions across all quizzes administered. In other words, c_z has a mean of 0 for all students and indicates whether, for a given question, they were more or less certain than was usual for their responses on other quizzes. Using this continuous variable, distributions could be compared in a more granular fashion. In particular, the relationship between correctness and certainty could be quantified using the common engineering tool of the Receiver Operating Characteristic curve, where the true positive (correct answer proportion) is plotted against the false positive (incorrect answer proportion) for each level of certainty. The AUC (Area Under the Curve) calculation was then used as a metric for the separation of the binary classes based on the normalized certainty axis. Other statistics for comparison between variables relied on chi-squared tests for independence or, for two binary categories, Fisher's exact test.

Results

Because they were only two questions long, students would often return completed quizzes to the Course Assistant before all quizzes had even been administered, achieving the "minute-quiz" goal. In the knowledge-testing questions, 14 items had a sample size over 10 (N between 14 and 18 responses); one question was discarded as being poorly aligned with learning outcomes and/or poorly worded (students had a response rate worse than would expected from guessing, and students reporting a higher certainty were more likely to answer incorrectly). Of the remaining 13 questions, a range of difficulties were present, from 6% to 94% correct. Despite this wide range, for 11 out of the 13 questions certainty was a reliable indicator of correctness as measured by the binary choice model of ROC analysis described in the methods section. The area under this curve (AUC) is related to how well the continuous variable of certainty discriminates between the binary outcomes of correct and incorrect. It is the probability that a randomly-chosen incorrect answer and randomly-chosen correct answer can be distinguished accurately solely on the basis of the one-dimensional certainty variable. Thus an AUC of 0.5 represents a truly random distribution of outcomes while an AUC of 1 corresponds to perfect correlation. A high AUC represents a highly accurate self-assessment on the part of students, both those that know they know the answer and those who know they don't know the answer; this will be referred to as "realistic" certainty. In contrast, low AUC represents "confusion" since students who were highly certain were incorrect and students who were correct were not certain. For six of the questions, AUC was greater than 0.8, indicating that knowing student certainty alone would be sufficient to tell the difference between a correct and incorrect answer more than 80% of the time. However, a high AUC does not necessarily mean that most students got the answer correct, as shown by the variety of responses in Figure 3.

The vertical axis is a normalized quantity of information (i_2) based on the proportion of students who got the answered correctly (p) and the expected proportion of correct answers if students were truly guessing (p_E) according to the following equation:

$$i_2 = \log_2 \left(\frac{1-p}{1-p_E} \right) \tag{1}$$

This is necessary because some questions were true-false, while other questions were multiple-choice or free-response. This means that some questions would have an expected correctness of 50% even if students were merely guessing. By normalizing against the information present in the options given, we obtain a better interpretation of the correctness proportion to represent how much extra information students had to inform their choice.

Based on the two variables in Figure 3, questions can be loosely divided into four quadrants with different interpretations. The upper right represents questions that students knew the answer to (high information) and for which they had highly accurate self-evaluation. These are "known knowns", a good indicator that the information in these questions was well-covered. The lower right represents questions that were difficult for students but which they understood; they were able to correctly assess whether or not they knew the answer. These might involve concepts the professor could review to improve retention, but there was no fundamental disconnect between question and answer. The questions were challenging but achieved the goal of prompting meaningful reflection. The third quadrant, however, is more interesting in its implications for

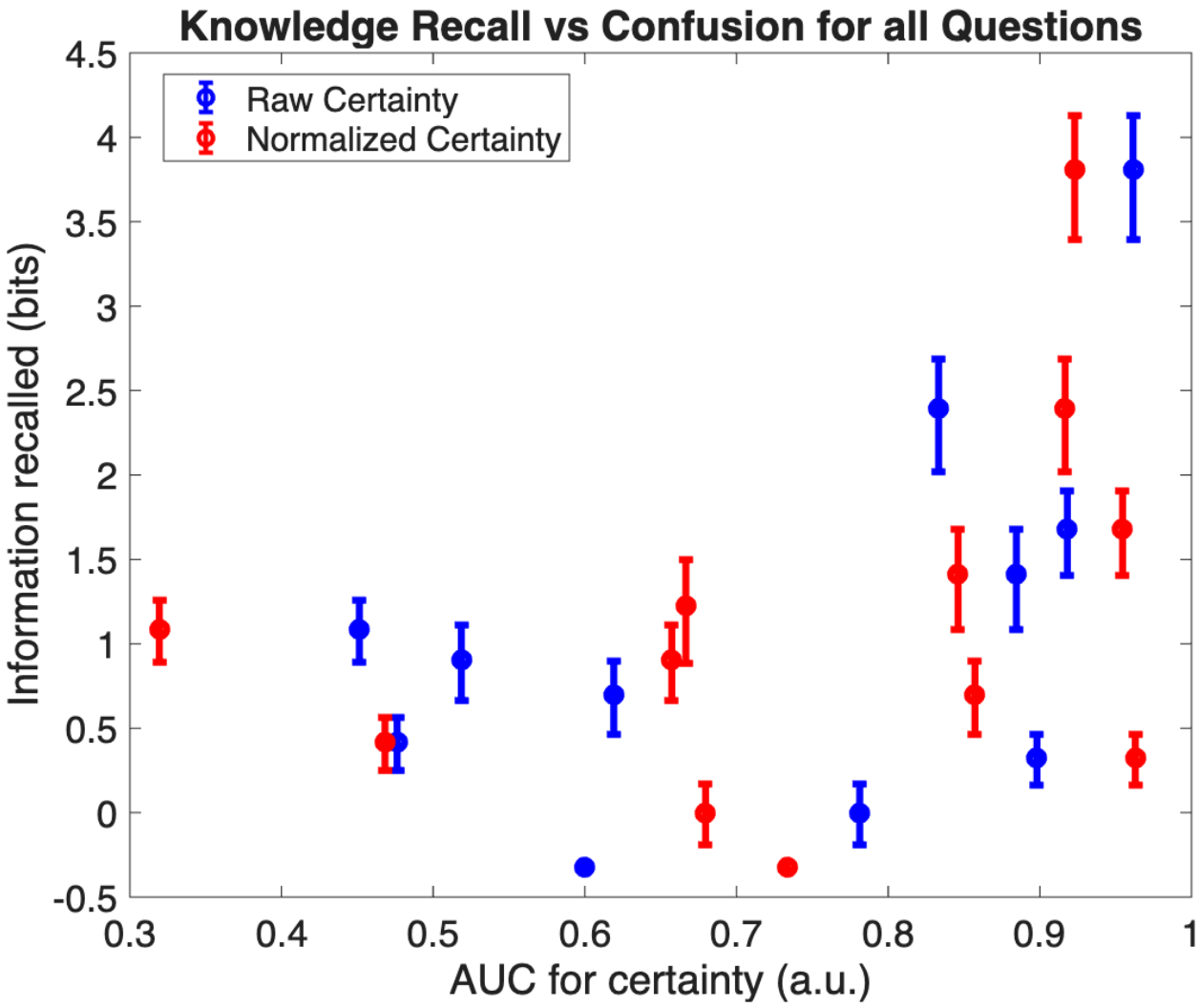

Figure 3: **Each question's aggregate information plotted against AUC.** Student knowledge recall, as described in equation 1, is a measure of the improvement over guessing which answer accuracy represents. The horizontal axis reports how realistic students' self-assessment of their certainty was, via the AUC metric. Each dot represents the aggregated results for one question, either evaluated using students' raw categorical certainty or normalized certainty based on individual trends. Error in information is based on the standard error of the mean for the empirical distribution.

question design. In the lower left, students experienced significant confusion and struggled to recall enough information to improve their answers. Perhaps students misunderstood the question in a way that caused them to over-estimate certainty. Or, perhaps the question was too wordy and thus looked more difficult than it was. A true-false item with too many clauses might also cause confusion since students aren't sure which part to focus on. These are questions that may be less useful since they caused students to incorrectly assess their own knowledge, and might need to be reworded or reevaluated to better align with learning outcomes. Finally, it is worth noting the absence of questions occupying the upper left of the graph: there were no questions that resulted in high confusion which students demonstrated a high information content, which is intuitively reasonable.

Observing the questions as a group, there are some emergent trends which warrant further study. First, as expected, responses marked with the highest certainty category were the most likely to be correct. As Wu et al. found, students who are willing to say they are completely certain of an answer are often over-estimating their accuracy (in their study, this threshold occurred at 85% certainty) [17]. This is roughly aligned with our finding that students who chose the highest level of certainty were still only correct in 86% of their responses (Figure 4A). At first glance, the middle two categories may seem redundant as they contained roughly the same distribution of correct and incorrect answers. However, examining one of the other independent variables in the data indicates that there was still a qualitative difference in how students perceived the "somewhat" and "mostly" certain categories. Namely, gender effects for certainty marking were present, with non-dominant gender students being overrepresented at low certainty and dominant-gender students overrepresented at high certainty (Figure 4B). This separation, when examined using the chi-squared test for independence, was significant at $p = .001$. This is despite the fact that the proportion of answers given correctly for each certainty level was not significantly different between the genders (the non-dominant gender students even tending to have slightly higher correctness at all levels of certainty; Figure 4C). It was also unsurprising to note that there was a strong effect of student major on reported certainty, with the Electrical Engineering students marking half of their responses as "Completely Certain". There were more male Electrical Engineering students than female, so to ensure that the observed gender difference was not simply an effect of student major, data were de-aggregated to only include Bioengineering students. The effect was still significant ($p = .007$) though not as extreme. These results indicate that it may still be useful to separate students by major in the analysis of next semester's data, since Electrical Engineering students tend to have a much higher base familiarity with the circuits concepts presented in the lecture sessions.

The previous analyses focused on students' categorical responses, a subjective measure based on their judgment at test-time. Using student-specific normalized certainty over the ensemble of quizzes (c_z), we can obtain a picture of the general behavior of each student and identify in which answers they were more certain than *usual*. This is our attempt to access a quantity representing the concept of "constructed certainty" in the conceptual framework model shown in Figure 1, though it is still inexact. Because c_z is a continuous variable, a more finely tuned analysis of correctness is possible, as shown in Figure 5. As expected, when students had higher certainty relative to other answers they had given, they were more likely to be correct; when they had low certainty, their correctness was no better than chance. To calculate each point on the curve, a moving window technique was employed. The central point, around a normalized certainty of c_z

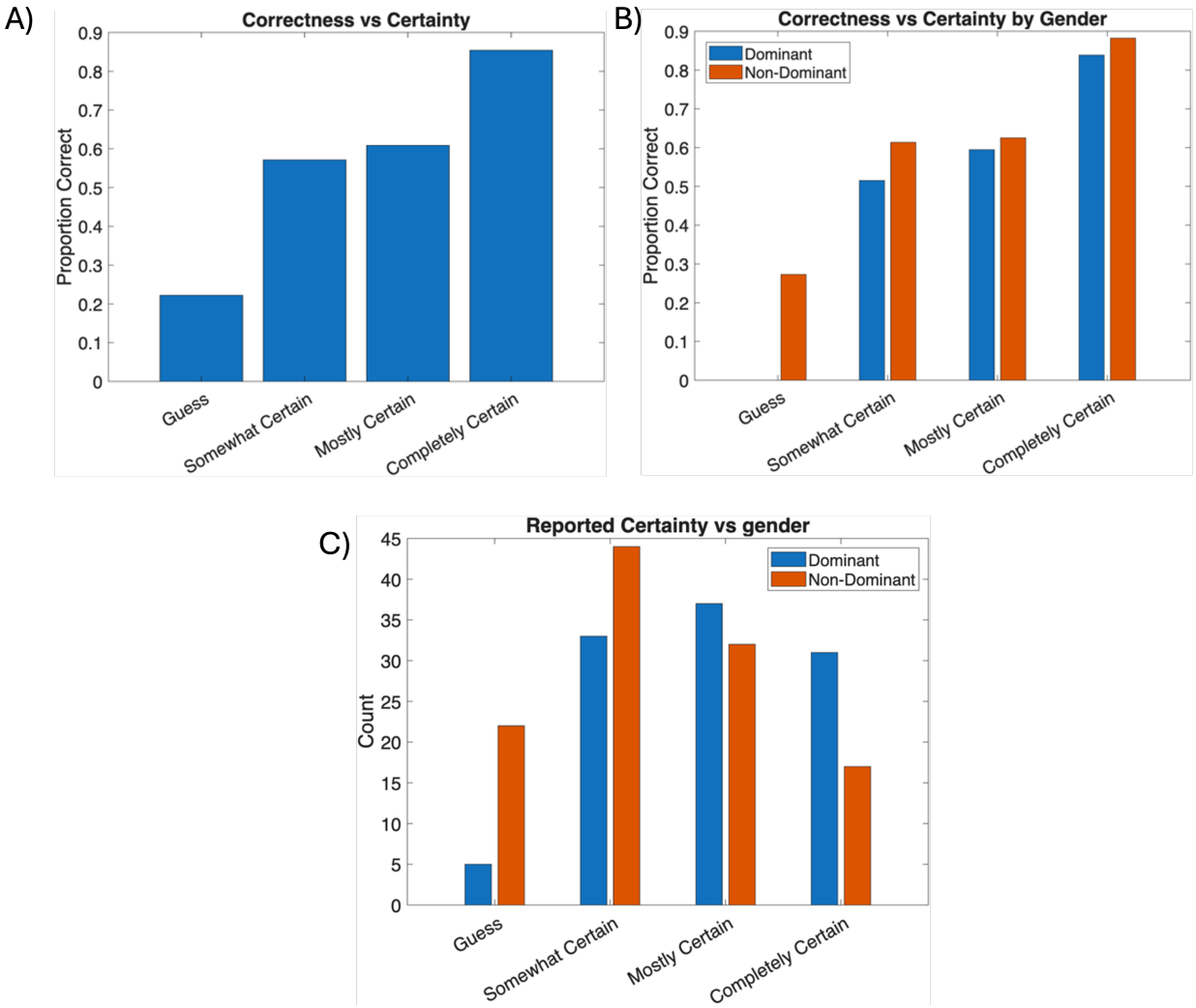

Figure 4: **Responses to knowledge-based questions according to categorical certainty and gender.** A) The average correctness increases with certainty level. B) De-aggregating by gender shows little difference within the categories that do not represent guessing. C) Trends between genders on giving responses in each category show a relative lack of reported certainty in the non-dominant gender students.

= 0, includes all answers with $-0.5 < c_z < 0.5$, of which approximately 60% were correct. The sample size, N, and the empirical correctness proportion p can be used to estimate the standard deviation of the expected correctness assuming a Bernoulli distribution of outcomes and using the following formula for the standard deviation of the mean of such a distribution:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}} \tag{2}$$

This is why the range of error grows significantly at the lower values of the reported c_z: relatively few answers were in the tail of the distribution with low certainty, but correctness was still near what would be expected from answers selected at random, resulting in a low denominator and increased numerator in eq. 4. To avoid overstating the precision of the error estimate, a moving-window average of width 1 normalized c_z unit was used to create a smoothed error representation. Finally, the second portion of each quiz contained "statement validation" questions in which each student was asked to rate a statement as being true or false, and their certainty. Different versions of the quiz presented the same statement as coming from different sources: either as the output of ChatGPT (a common LLM) or as the contribution of a peer when completing an assignment together. Some of the statements were intentionally worded in an ambiguous manner to see if students would react with more skepticism; this did not have a significant effect on responses for either source or in aggregate. For the subjective statements, more students marked them as being "possibly misleading" when ChatGPT was the source, but this difference was not significant; more data will be needed to examine this trend. However, source did have an impact on whether students were willing to agree with objective statements that were definitively true or false. Specifically, some students were more reluctant to affirm a statement as true if ChatGPT was the source (p = 0.096 by the Fisher exact test), while false statements were evaluated roughly equivalently regardless of provenance (Figure 6). Finally, paralleling the results from the knowledge check questions, the trend of gender impacting the prevalence of low- and high-certainty evaluations was present in these data as well (p = 0.02).

Discussion

The results demonstrate that asking students to consider their certainty when answering can provide valuable insights about how questions are received and whether students have the depth of knowledge required to accurately claim ignorance. For most questions, students were realistically assessing their certainty in their answer, which is a sign of good metacognition and critical thinking. However, examining the questions for which this is not the case is just as illustrative, since it has implications for question design. By using this type of certainty analysis, three independent axes of question quality are available: difficulty, or the proportion of correct answers; average certainty, or how difficult the question was perceived to be by students as they completed it; and whether the population of student certainty is realistic, which is calculated by how well an elevated certainty correlated with a correct response. This means that questions can be separated into categories such as "known unknowns" in which students accurately assess themselves as uncertain when they tackle a tough question, or "misconceptions" when students are highly confident but that confidence does not translate to correctness. By examining these outputs, an educator may be able to assess whether the content tested in such a way needs to be reviewed, explained differently, or made more challenging.

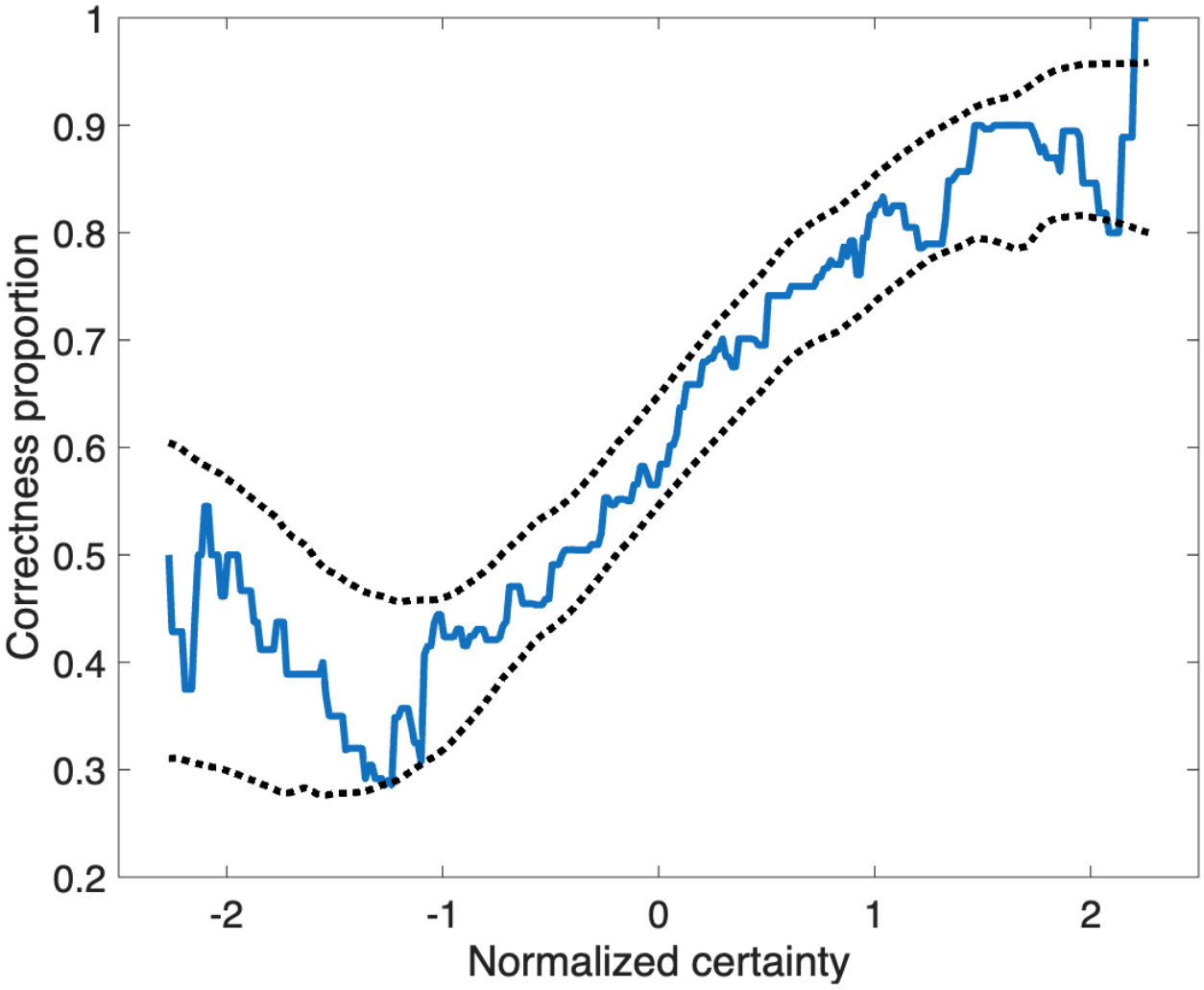

Figure 5: **Student correctness vs normalized certainty score for all responses.** The central line represents the proportion of answers that were correct for the population of responses within a window of one standard deviation surrounding the certainty level on the horizontal axis. The dotted lines are a smoothed representation of the standard deviation of the mean correctness for each level (eq. 2).

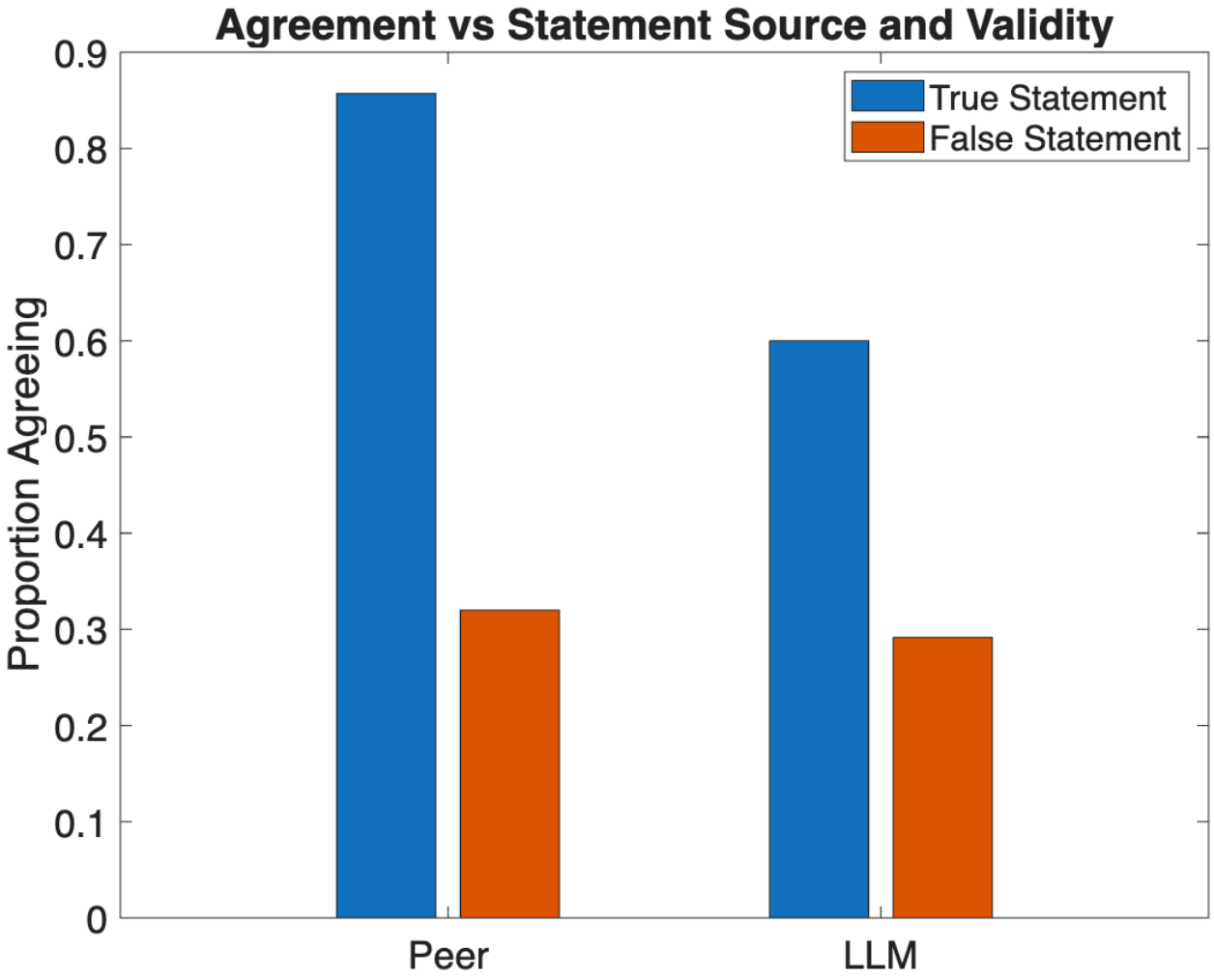

Figure 6: **Affirmation of statement truth depending on source and validity.** As expected, when presented with a false statement students were more likely to disagree, and this trend was persistent regardless of statement source. However, when a true statement was presented as authored by an LLM there was a tendency to doubt, more than when the same statement came from a peer. This difference is approaching significance via the Fisher exact test ($p = 0.096$).

Although large studies on objective probability-based certainty do not show significant differences between students' performance based on gender, we hypothesized that this group of students who were not accustomed to self-reporting certainty may exhibit systematic differences when choosing the qualitative statement that best describes their self-assessment. This is based on prior work on a similar population of students when assessing their expected self-efficacy in various engineering-related tasks [23]. As hypothesized, male students tended to be overrepresented at higher levels of certainty despite small differences in correctness, while non-male students were overrepresented in the expression of acknowledged ignorance. Interpreting this result is complex given the limited data available. Students provided consent to use their data, so some self-selection is present in the resulting dataset. In addition, it is not possible to separate the impacts of personality and the impacts of perceived expectation as influences on student behavior. Regardless, this is an important demonstration of why categorical statements of certainty are inappropriate for assigning grades in a summative assessment. The formative assessment tool here is designed as an aid to student self-reflection, not as a means of quantitatively rewarding students based on realistic certainty levels. This formative nature does potentially compromise the study of student knowledge in this setting, as there is no extrinsic motivation for students to perform thoughtfully on this assessment. It would therefore be necessary to incorporate this assessment type into a classroom in a collaborative manner where students understand the significance of ungraded work, or where another incentive is given to encourage students to answer to the best of their ability.

Finally, there was some evidence that students consider the source of information when they evaluate the truth of a statement, but this may be limited to a mild skepticism for ChatGPT output. This difference appeared to come from particular students skewing the data, which indicates that a larger sample size may allow for clearer partitioning of the class into "LLM trusting" and "LLM skeptical" groups. Some mistrust may be healthy when considering technical information in LLM output, but it's important to note that sources of varying trustworthiness come in many forms. The goal is not to create a specific attitude towards any one source but rather to foster a prudent self-evaluation of knowledge whenever students find themselves searching for explanations. If they have the background to accurately assess their certainty, it will give them the ability to take accountability for information they cite—and, more importantly, to admit when they aren't sure and look for corroborating sources.

Conclusions and Future Work

This study shows that even simple formative assessments can be a valuable place to encourage metacognition. By including a reported certainty on basic knowledge recall questions, quiz items become informative for the instructor beyond the raw percentage correct. Students in this study occasionally exhibited overconfidence, an effect that depended on both gender and major. And there was some suggestion that LLMs as a source resulted in a different mental evaluation of truthfulness. The quantitative data are suggestive for future work, but they also suggest a path forward for expanding the presence of self-reflection in technical curriculum. This study seeks to convince educators that certainty is a valuable skill for engineers to possess which can be incorporated into their assessment structure with little additional effort. Students who learn when they can trust their memory and judgment are more equipped to navigate a world with an

increasing number of knowledge sources of varying quality. Those who gain familiarity with acknowledging ignorance in an assessment scenario may feel more comfortable admitting uncertainty in their careers when their decisions will carry dire consequences for false confidence.

The lessons learned in the first semester will be used to improve the study for implementation in the spring. Considerably more students are enrolled, which will allow for better determination of the significance of small effects. Most questions will be kept the same, but those which were determined to be poorly aligned with course outcomes will be updated. This study could be expanded in the future by incorporating a student feedback survey to determine whether students felt they had improved their metacognitive skills, or whether they regularly use LLMs. Finally, incorporating the question format and analysis methods used in this study into a learning management system would be the ultimate test of the feasibility for wide-spread adoption of formative certainty assessment.

References

[1] M. Ghassemi, A. Birhane, M. Bilal, S. Kankaria, C. Malone, E. Mollick, and F. Tustumi, "ChatGPT one year on: who is using it, how and why?" *Nature*, vol. 624, no. 7990, pp. 39–41, Dec. 2023. doi: 10.1038/d41586-023-03798-6 .

[2] A. Extance, "ChatGPT has entered the classroom: how LLMs could transform education," *Nature*, vol. 623, no. 7987, pp. 474–477, Nov. 2023. doi: 10.1038/d41586-023-03507-3 .

[3] B. Diaz, G. Chen, E. Jaselskis, and C. Delgado, "Supporting Generative AI Literacy: Exploring the Pedagogical Roles Students Assign ChatGPT and Impact on Course Grades," *Comunicar*, vol. 33, no. 82, pp. 46–61, Jul. 2025. doi: 10.5281/zenodo.15993999 .

[4] M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, L. W. Mayer, and P. Smyth, "What large language models know and what people think they know," *Nature Machine Intelligence*, vol. 7, no. 2, pp. 221–231, 2025. doi: 10.1038/s42256-024-00976-7 .

[5] G. Puthumanaillam, T. Bretl, and M. Ornik, "The Lazy Student's Dream: ChatGPT Passing an Engineering Course on Its Own," *IFAC-PapersOnLine*, vol. 59, no. 7, pp. 213–218, Jan. 2025. doi: 10.1016/j.ifacol.2025.08.049 .

[6] M. Preheim, J. Dorfmeister, and E. Snow, "Assessing Confidence and Certainty of Students in an Undergraduate Linear Algebra Course," *Journal for STEM Education Research*, vol. 6, no. 1, pp. 159–180, 2023. doi: 10.1007/s41979-022-00082-6 .

[7] L. W. Anderson and D. R. Krathwohl, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

[8] H. O. Soderquist, "A New Method of Weighting Scores in a True-False Test," *The Journal of Educational Research*, vol. 30, no. 4, pp. 290–292, 1936. [Online]. Available: https://www.jstor.org/stable/27526229

[9] S. Plous, *The psychology of judgment and decision making*. Mcgraw-Hill Book Company, 1993.

[10] S. S. Jacobs, "Correlates of Unwarranted Confidence in Responses to Objective Test Items," *Journal of Educational Measurement*, vol. 8, no. 1, pp. 15–20, 1971. [Online]. Available: https://www.jstor.org/stable/1434247

[11] N. Kogan and M. A. Wallach, *Risk taking: A study in cognition and personality*, ser. Risk taking: A study in cognition and personality. Oxford, England: Holt, Rinehart & Winston, 1964.

[12] R. Hansen, "The Influence of Variables Other than Knowledge on Probabilistic Tests," *Journal of Educational Measurement*, vol. 8, no. 1, pp. 9–14, 1971. [Online]. Available: https://www.jstor.org/stable/1434246

[13] G. J. Echternacht, R. F. Boldt, and W. S. Sellman, "Personality Influences on Confidence Test Scores," *Journal of Educational Measurement*, vol. 9, no. 3, pp. 235–241, 1972. [Online]. Available: https://www.jstor.org/stable/1434170

[14] X. Xu, S. Kauer, and S. Tupy, "Multiple-choice questions: Tips for optimizing assessment in-seat and online," *Scholarship of Teaching and Learning in Psychology*, vol. 2, no. 2, pp. 147–158, 2016. doi: 10.1037/stl0000062 .

[15] A. R. Gardner-Medwin and M. Gahan, "Formative and summative confidence-based assessment," in *Proceedings for 7th CAA Conference 2003*. Leicestershire, UK: Loughborough University, 2003. [Online]. Available: https://caaconference.lboro.ac.uk/pastConferences/2003/procedings/gardner-medwin.pdf

[16] P. Hassmén and D. P. Hunt, "Human Self-Assessment in Multiple-Choice Testing," *Journal of Educational Measurement*, vol. 31, no. 2, pp. 149–160, 1994. [Online]. Available: https://www.jstor.org/stable/1435174

[17] Q. Wu, M. Vanerum, A. Agten, A. Christiansen, F. Vandenabeele, J. M. Rigo, and R. Janssen, "Certainty-Based Marking on Multiple-Choice Items: Psychometrics Meets Decision Theory," *Psychometrika*, vol. 86, no. 2, pp. 518–543, Jun. 2021. doi: 10.1007/s11336-021-09759-0 .

[18] A. R. Gardner-Medwin, "Confidence-based marking: towards deeper learning and better exams," in *Innovative assessment in higher education*. Routledge, 2006, pp. 161–169.

[19] T. Gardner-Medwin and N. Curtin, "Certainty-based marking (CBM) for reflective learning and proper knowledge assessment," in *REAP International Online Conference on Assessment Design for Learner Responsibility*, 2007, pp. 1–7. [Online]. Available: https://tmedwin.net/~ucgbarg/tea/REAP/REAP_cbm.pdf

[20] D. A. Barr and J. R. Burke, "Using confidence-based marking in a laboratory setting: A tool for student self-assessment and learning," *J Chiropr Educ*, vol. 27, no. 1, pp. 21–6, 2013. doi: 10.7899/JCE-12-018 .

[21] G. Yuen-Reed and K. B. Reed, "Engineering Student Self-Assessment through Confidence-Based Scoring," *Advances in Engineering Education*, vol. 4, no. 4, 2015. [Online]. Available: https://advances.asee.org/wp-content/uploads/vol04/issue04/Papers/AEE-16-Reed.pdf

[22] C. Clark, "The impact of confidence-based marking on unit exam achievement in a high school physical science course," Graduate Research Papers, University of Northern Iowa, 2020. [Online]. Available: https://scholarworks.uni.edu/cgi/viewcontent.cgi?article=2448&context=grp

[23] H. R. C. Kimmel, M. Agrawal, J. Tibbs, K. Tuvilleja, and R. M. Reck, "Work In Progress: Adding Additional Methods to Identify Mistakes in an Undergraduate Biomedical Instrumentation Laboratory Course," in *2025 ASEE Annual Conference & Exposition*, Jun. 2025. [Online]. Available: https://peer.asee.org/57475

[24] V. Kovanović, A. Barthakur, S. Joksimović, and G. Siemens, "Why universities need to radically rethink exams in the age of AI," *Nature.*, vol. 648, no. 8092, pp. 35–37, 2025. doi: 10.1038/d41586-025-03915-7 .

[25] R. Dennick, "Twelve tips for incorporating educational theory into teaching practices," *Med Teach*, vol. 34, no. 8, pp. 618–24, 2012. doi: 10.3109/0142159X.2012.668244 .