2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
ASEE
Paper ID #39790

# Concept-Centric Summative Assessments That Remain Authentic while Reducing Grading Effort

**Prof. Curt Schurgers, University of California, San Diego**

Curt Schurgers is a Teaching Professor in the UCSD Electrical and Computer Engineering Department. His research and teaching are focused on course redesign, active learning, and project-based learning. He also co-directs a hands-on undergraduate research program called Engineers for Exploration, in which students apply their engineering knowledge to problems in exploration and conservation.

# Concept-Centric Summative Assessments that Remain Authentic while Reducing Grading Effort


## Abstract

Engineering courses often evaluate students' comprehension by having them solve problems as part of their summative assessments. However, grading these assessments based on students' worked solutions can be time-consuming and labor-intensive. To overcome this challenge, this paper proposes a design methodology that reduces grading time while maintaining authenticity, by using a question design strategy that allows grading to be based on final answers only. The effectiveness of this approach was validated in a large undergraduate electrical engineering course, where students lost less than 5% of partial credit on average, while the time required for grading was reduced eight-fold.


## Introduction

Learning outcomes of engineering courses typically involve students being able to understand and apply concepts, rather than just memorize or reproduce information. To evaluate whether these learning outcomes have been met, authentic summative assessments (i.e., summative assessments that evaluate whether students can successfully transfer the knowledge and skills [26]) focus on students applying the concepts that they have learned. This means that these assessments often involve students working through problems. For example, in the lower division electrical engineering course used to validate this study, the final exam takes 3 hours and normally consists of 5 to 6 such constructed-response problems.

Generally, these problem-based summative assessments involve a holistic grading approach, based on the fully worked solution. They are not well-suited for a format where grades would be based on the final answer only, as this limits the opportunity to award partial credit. Questions often require students to successfully apply multiple concepts embedded in the same problem. Instead of considering the final answer only, the students' problem -solving strategy itself is considered in detail, by evaluating the steps they follow to arrive at the answer. In the literature, this type of exam that assigns grades based a student's worked response is referred to as "constructed-response" (CR) [8]. Grading rubrics are used to capture the richness of the problem-solving approach, to arrive at a more authentic assessment.

The challenge is that this approach requires significant grading effort and time. The goal of our study is to investigate a design strategy for summative assessments that significantly cuts down on grading time, while keeping the assessment authentic. A reduction in grading time would free up resources that could be redeployed in other parts of the course, such as providing more

tutoring support. In this paper, we propose a design methodology to create such time-efficient authentic summative assessments for engineering courses. Our proposed design strategy consists of two steps: (1) systematically create a problem that consists of targeted sub-questions and (2) grade these based on a rubric applied to the final answers only. We will validate the representative nature of this summative assessment methodology by implementing it for a large undergraduate electrical engineering course with an enrollment of 275+ students. By using a comparison group, we will evaluate how close the new approach comes to capturing the richness of the original problem-based assessment.

## Related Work

### *Multiple-choice exams*

Reduction of grading effort has received significant attention in higher education. A strategy that has been well-established in this regard is the use of multiple-choice questions [1-14]. It has important similarities to our proposed approach. However, the fact that a selection of answer choices is available to students in multiple-choice exams, whereas it is not in our approach, is an important distinction, which we will comment on in detail. Nevertheless, the literature on multiple-choice exams provides important insights.

Advantages of multiple-choice exams from the perspective of the instructor are the ability to easily create multiple versions, increased question-granularity that allows many topics to be covered, and ease of automated grading [3]. Students, on the other hand, see advantages such as lower instructor grading bias, the ability for partial credit due to more questions being asked, the perception of the exams being easier, and the related fact that answers may be guessed [3, 4]. For these reasons, some research has reported students often preferring multiple-choice exams [1, 9]. Other research presents a more balanced picture in terms of preference over constructed-response exams, where students falling in one camp versus the other is attributed to their opposing views on answer guessing, selecting versus constructing answers, and perceived ability to demonstrate knowledge [8].

A key question is that of validity of multiple-choice exams in measuring performance. Educational psychology demonstrates that it is theoretically possible to construct multiple-choice questions that measure many of the same cognitive abilities as constructed response ones [5, 6]. However, these studies also stress the need for empirical testing and validation. In this regard, existing work has reported a range of results, with some concluding that they are able to assess the same knowledge as constructed response questions [12, 13], while others concluding the opposite [6, 11]. Studies have shown that the reason for these differencing results may be caused by significant differences in the design of the multiple-choice questions [3]. These different designs may test very different levels of student understanding, from superficial knowledge and

rote memorization to deeper levels of understanding (e.g., when they implicitly require problem solving) [16]. Current research has consequently pointed out that careful design can overcome some of the issues commonly associated with multiple-choice exams [3]. This includes efforts to reduce the impact of guessing, such as through more advanced approaches to how points are awarded [2, 7, 15].

A key observation remains the importance of good question design [3]. Furthermore, studies have also found evidence that anticipatory learning, i.e., studying differently depending on the type of test students anticipate, is a factor in considering the validity of multiple-choice exams [5]. If students perceive multiple-choice exams as easier, susceptible to guessing or only requiring memorization [1, 9], they will fail to move away from memorization and superficial learning towards deeper learning [10]. Since prewritten alternatives are provided, even perceived impact of guessing may thus impact effectiveness [3, 4]. This is an important distinction with the proposed approach in this paper, which remains open-ended, i.e., it does not rely on provided answer choices.

***Automated grading***

An important reason for instructors to consider multiple-choice tests is the ease with which they can be automated [3]. As a more general trend, digital and web-based tools that streamline the grading process, such as Gradescope [21], have shown to yield advantages in overall workflow, reliability, consistency, as well as grading efficiency, even without grading automation [17]. As these platforms become increasingly sophisticated, these tools are also offering grade automation for more than only multiple-choice questions. Advantages of this automation are not only improved scalability, but also a reported reduction in susceptibility to grader error and inconsistency [25]. For example, platforms such as Gradescope, have incorporated AI (artificial intelligence) into their grading tool suite [19]. This includes handwriting recognition, such that scanned responses can be mapped to their digital equivalent. Instructors upload templates to indicate where the AI tool can find specific information on the scanned pages (e.g., areas where students are required to write their name or final answers to a question). This is useful to automatically assign exams to the appropriate student, by reading their name and/or identifier. Also, it allows decoding final answers and creating groups of identical answers, which can then be graded jointly at the group-level. We will leverage this feature in our implementation.

The need to scale the grading process, and the resulting use of automation, is also key to Massive Open Online Courses (MOOCs) due to their large enrollments. MOOCs rely on a variety of grading approaches, with the more open-ended approaches considered more valuable. These approaches include multiple-choice, checking a numerical value or symbol, fill-in-the-blank derivations, drag-and-drop drawing, and short-answer style quizzes [23]. The latter, in which students are asked to provide a short free-response answer are often peer- or self-graded, due to difficulty in automation. However, research is conducted towards the use of AI there as well [20,

22]. Most relevant to our work are the fill-in-the-blank derivations, as they relate most closely to evaluating process and reasoning. In this answer format, students are provided answer boxes at multiple steps within a longer derivation. However, while valuable in MOOCs, the effort involved with creating these kinds of questions does not scale for one-off exams in traditional course settings.

Most relevant to our study is the recent work by Veale and Craig on final answer assessment in a linear algebra course, which was also motivated by a reduction in grading effort afforded by automation [24]. They focus on the same approach as we do -- strictly relying on final answers to assign grades in constructed response questions. The focus of their work is addressing the important question of validity. In mathematics, as in engineering, the way in which one arrives at the answer is a key assessment metric. Their validation consists in grading constructed response questions two-ways: based on final answers only and based on the fully worked solution. We will follow the same approach. Their main contribution is to present a set of design principles that constructed-response questions should satisfy to improve their validity when graded based on final answers only. These principles are (1) to limit the number of points per subquestion, (2) to focus on problems that are minimally susceptible to careless error or are quickly checkable, (3) to limit the algebra involved and test the core concepts, (4) to avoid the same error being penalized twice, and (5) to make sure to check all possible answers. This ties into similar observations for multiple-choice exams, which claim that validity can be largely overcome by careful question design [3].

We will build upon the five design principles by Veale and Craig. However, the authors do not provide guidance on ***how*** to design questions according to these design principles. Our contribution is specifically to address this question, i.e., ***to propose a methodology to systematically design high-validity questions.***

## Methods

### *Design Methodology*

Our proposed methodology is to leverage the grading rubric of the traditional problem-based version of the summative assessment, to isolate the core concepts that are being tested. These core concepts then serve as the building blocks to create new targeted mini-questions. In effect, a traditional problem is modified this way to be made up of a collection of subparts instead. These conceptual-building-block subparts (subquestions) are then subsequently graded by only evaluating the final answer. This two-step methodology is further detailed below:

**Step 1 - Problem construction**

a. Design a traditional problem-based constructed-response problem, which tests a set of concepts within a single problem writeup. We will refer to it as a "***holistic***" design (to contrast it to our new design).
b. Create the grading rubric for this holistic problem. Considering where one assigns partial credit will pull out the core concepts the problem is assessing. If necessary, break these down further until each carries the same weight in the grading rubric.
c. Translate each equally-weighted rubric item into a dedicated subquestion. Each subquestion can be an element of a larger question (e.g., a different quantity that students must calculate for the same circuit). However, it is important that these subquestions are as independent as possible, to avoid a single mistake propagating through [24]. We will refer to our new problem design, consisting of these subquestions, as a ***building-block design***.

**Step 2 – Rubric creation**

Create a new grading rubric to be applied to the final answers of the building-block design. This rubric can be more granular than just correct/incorrect (note that related work did not use such a rubric and only graded based on binary correctness [24]). For example, we could:

- assign half credit if the answer is correct except for a sign error (assuming that this would indicate at least partial mastery).
- assign full or half credit if comparing to a prior answer can yield insight. For example, a concept could relate to the expected change or non-change of a quantity due to a change in the circuit. Two subproblems could ask for the value before and after the change respectively, with full credit awarded for the second part if the difference is correct rather than the value itself. This approach is useful when a concept is more important and needs to be split into sub-concepts in Step 1, to achieve equal weighting.

***Example***

We will illustrate the proposed approach through an application example. Figure 1 shows the holistic exam question that served as the starting point of the procedure (Step 1). By creating the grading rubric for this question, shown in Table 1, this holistic exam version was analyzed as testing students' knowledge of three key concepts:

(1) Solving the circuit for $V_a$, with the challenge of a dependent source in the circuit.
(2) Calculating the average power $P_s$ based on the complex power $S$ [18].
(3) Solving for $S$ symbolically first, which aids in discovering that the source's phase does not impact the final result.

The circuit below represents an AC circuit in steady-state in the phasor domain (for the complex numbers, you may assume units are V, A, Ω, etc. as appropriate). The current source $i_S$ is an AC source with $\omega = 5$ rad/s. Each box represents the impedance of a single circuit element (a resistor, capacitor or inductor). Find the average power $P_s$ supplied by the voltage source.

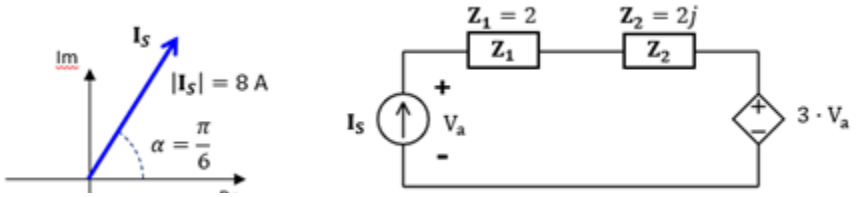


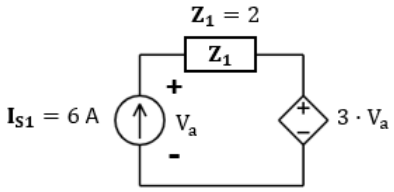

**Figure 1.** Example of a holistic problem design.

**Table 1.** Grading rubric for the holistic problem example.

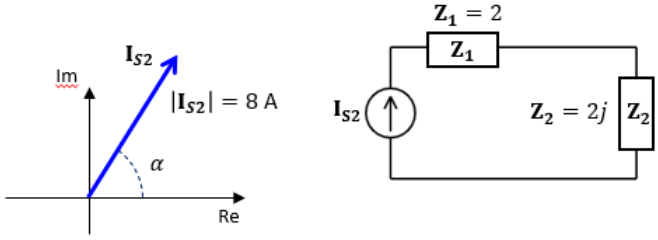
| | Holistic problem design | |
|---|---|---|
| Entire problem | -0.5 | Calculation error |
| | -0.5 | Sign error power supplied/received |
| | -0.5 | Answer with complex power instead of average power |
| | -1 | Error finding equation for $V_a$ |
| | -0.5 | Solving it numerically and making a small calculation error |
| | -1.5 | Solving it numerically and making a significant error |
| | -3 | Incorrect approach/blank |

The circuits below represent AC circuits in steady-state in the phasor domain (for the complex numbers, you may assume units are V, A, Ω, etc. as appropriate). The independent current sources are AC sources with $\omega = 5$ rad/s. Each box represents the impedance of a single circuit element (a resistor, capacitor or inductor).

(a) Find the phasor $V_a$. You can write your answer in cartesian or polar coordinates.




(b) In the circuit below, we set $\alpha = \frac{\pi}{2}$ (the figure is clearly not drawn to scale). Find the average power $P_s$ <u>supplied</u> by the current source $I_{s2}$.




(c) For the same circuit as in part (b), we change $\alpha$ to $\alpha = \frac{\pi}{6}$. Find the average power $P_s$ <u>supplied</u> by the current source $I_{s2}$.

**Figure 2.** Example of a building-block problem design.

These three concepts then gave rise to the three subproblems that made up the new question in the building-block design, shown in Figure 2.

Next, the new grading rubric is created for the building-problem design, corresponding to Step 2 of our methodology. The resulting rubric is shown in the right-most column of Table 2, labeled as ***correctness grading***. This correctness grading is the rubric that is the core of the proposed approach: only consider the final answer. It also demonstrates three examples of grading granularity: part (a) is purely binary (correct/incorrect), part (b) allows also for half-credit in a way that is easily checked (opposite sign) and part (c) checks correctness by comparing to the answer in part b (full credit is awarded if the answer is the same).

Table 2 also includes an alternative rubric, which we call *comprehensive grading* (it is called "hypothetical" grading in related work [24]; we avoid this terminology to limit confusion). This is a rubric that one would apply if one were to grade the building-block design (Figure 2) in the traditional way, i.e., by considering the student's worked solution rather than only the final answer. The reason to include this approach in our study is that it allows us to evaluate the impact of each of the two steps in our methodology individually. By comparing the original holistic design (Figure 1 with the rubric of Table 1) to the building-block design with comprehensive grading, we can isolate the effect of redesigning the problem into a collection of subproblems (Step 1), i.e., *the problem design*. On the other hand, by comparing the correctness grading versus the comprehensive grading for the same building-block design, we can evaluate the impact of considering only the final answer (Step 2), i.e., *the grading approach*.

**Table 2.** Grading rubric for the building-block problem example.

| | Building-block problem design | | |
|---|---|---|---|
| | Comprehensive grading | | Correctness grading |
| Part (a) | -0.5 | Calculation error | -1 |
| | -1 | Error finding equation for $V_a$ | |
| Part (b) | -0.5 | Error finding equation for $V_{across}$ | -0.5 if sign error, else -1 |
| | -0.5 | Calculation error | |
| | -0.5 | Sign error power supplied/received | |
| | -0.5 | Answer with complex power instead of average power | |
| | -0.75 | Solving it numerically and making an error | |
| | -1 | Incorrect approach/blank | |
| Part (c) | -0.25 | No explanation or calculation for answer | 0 if equal to (b), else -1 |
| | -0.5 | Recalculating and making a calculation error | |
| | -1 | Recalculating and making a major error | |

## Experiment Setup

To validate our proposed methodology, we implemented it for the final exam of ECE 35, an introduction to electrical circuits course at UC San Diego, in Fall'22. Due to its large enrollment, the class was split into two sections: Section A with 151 students and Section B with 128 students. Both sections were offered by the same instructor. The final exam was a 3-hour written test, on Tuesday of finals week for Section A and on Thursday for Section B. As with other written tests in the course, students were not allowed the use of calculators. Instead, all questions were designed to only involve numbers and calculations that were easy to do by hand. The final exam for each section consisted of 5 questions. First, two exams were created, both in the traditional holistic style, which we refer to as *E1-H* and *E2-H*. The corresponding questions on these two exams were distinct but tested on the same course topics. Next, one of the exams, *E1-H*, was converted into a building-block design, which we refer to as *E1-BB*, using the 2-step approach we described earlier. This resulted in one holistic problem (from *E2-H*) and one building-block problem (from *E1-BB*) on each of the five course topics. The questions of these two versions were then distributed across the finals for the two sections as shown in Table 3. Note that the example shown in Figure 2 was part of *E1-BB* and used as the first part of Question 4 for Section B. For brevity, the other half of that exam question was not shown. Figure 1 showed the corresponding *E1-H* version of that part of the problem (note that the *E1-H* questions were only used to generate the *E1-BB* version and not part of the final, see also Table 3).

**Table 3.** Exam problem assignments for the two sections.

| | Section A (N = 151) | Section B (N = 128) |
|---|---|---|
| Question 1 (5 points) | Building-block design (*E1-BB*) | Holistic design (*E2-H*) |
| Question 2 (7 points) | Building-block design (*E1-BB*) | Holistic design (*E2-H*) |
| Question 3 (6 points) | Holistic design (*E2-H*) | Building-block design (*E1-BB*) |
| Question 4 (6 points) | Holistic design (*E2-H*) | Building-block design (*E1-BB*) |
| Question 5 (7 points) | Holistic design (*E2-H*) | Building-block design (*E1-BB*) |

Regardless of the version of the problem they got, students submitted their worked solution (i.e., the derivations, analysis, etc.) and also wrote their final answer in a dedicated answer box. If they had the holistic version of a problem, it was graded according to its corresponding rubric applied to the worked solution. If the students got the building-block version of the question, it was graded using the comprehensive rubric, again based on the worked solution. This was done because the proposed methodology that uses correctness grading of the final answer only, was still under study. As such, to ensure fairness and consistency of course grades, the student's work was always considered, whether they received the holistic design and the building-block design of a problem. The splitting of the question types across the two sections was done to ensure fairness across the student pool in case there was an inherent difference in the two designs. Note that the exams of the two sections originated from two distinct versions, and therefore the exam that each section received was sufficiently different to avoid issues with academic integrity. For

each question, the holistic design resulted in a comparison group for the building-block design. For the purpose of this study, correctness grading was then also applied to the building-block design. This approach is the core of our proposed methodology. An AI-assisted grader was used to facilitate the correctness-based grading methodology. Specifically, exams were uploaded into Gradescope [19]. Its built-in AI was used to read the handwritten answers from the answer boxes and automatically classify them in groups. These groups were then assigned grades in bulk. Answers that the AI could not process were kept separate for manual classification.

## Results

### *Grading Approach Comparison*

First, we investigate the impact of the grading approach by itself (Step 2). We do this by comparing correctness and comprehensive grading for the building-block design. Figure 3 on the next page shows the grades with one strategy versus the other. The area of the bubble is proportional to the number of students represented by that data point. Note that Question 1 and 2 were given to section A (N = 151), while the other three questions were given to section B (N = 128), see Table 3 from before. The dotted red lines in the figure represent the cases when the two scores match perfectly or are off by only one point in either direction. This allows us to compare the actual grades to the ideal case, i.e., where all the bubble centers lie on the center red line.

In investigating Figure 3, we observe that for some questions (such as Questions 4 and 5), the correctness grading is a closer match to the comprehensive grading. To get a better insight into why the discrepancy varies, we plotted the histograms of the score differences in Figure 4 on the following page. The score difference is defined as the score under comprehensive grading minus that under correctness grading. Positive values represent partial credit resulting from the comprehensive grading approach that considers the worked solution. On the other hand, negative values occur in situations where students received more points with correctness grading. This happened in two scenarios: (1) students correctly guessed the final answer without any supporting work (or forgot to write down their work or reasoning), or (2) students accidentally got the correct final even though they used an incorrect approach. In both these cases, the grader looking at the worked solution would not award points, even though the students had the correct final answer and thus received the points when correctness grading was used.

In Figure 4, the different shades of green correspond to the number of partial credit instances under comprehensive grading, i.e., the number of subparts of the question that were awarded partial credit. For example, a bar in pale green corresponding to “3 instances of extra credit” with a score difference of 1.25 means that students received partial credit on three subparts of the problem. The yellow bar corresponds to no discrepancy between two grading schemes (0 difference). This is the most common case.

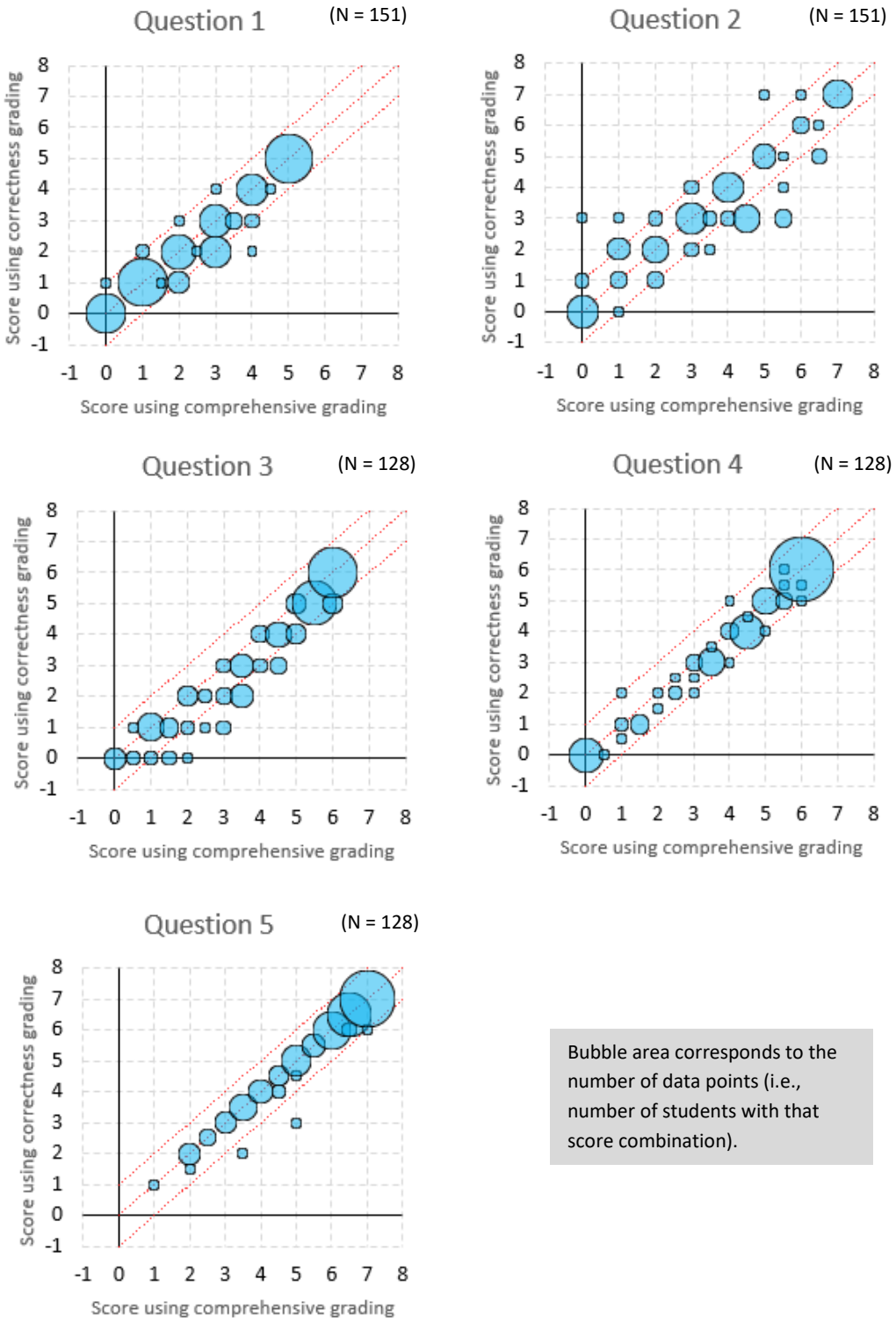



**Figure 3.** Comparison of comprehensive and correctness grading (building-block problem design).

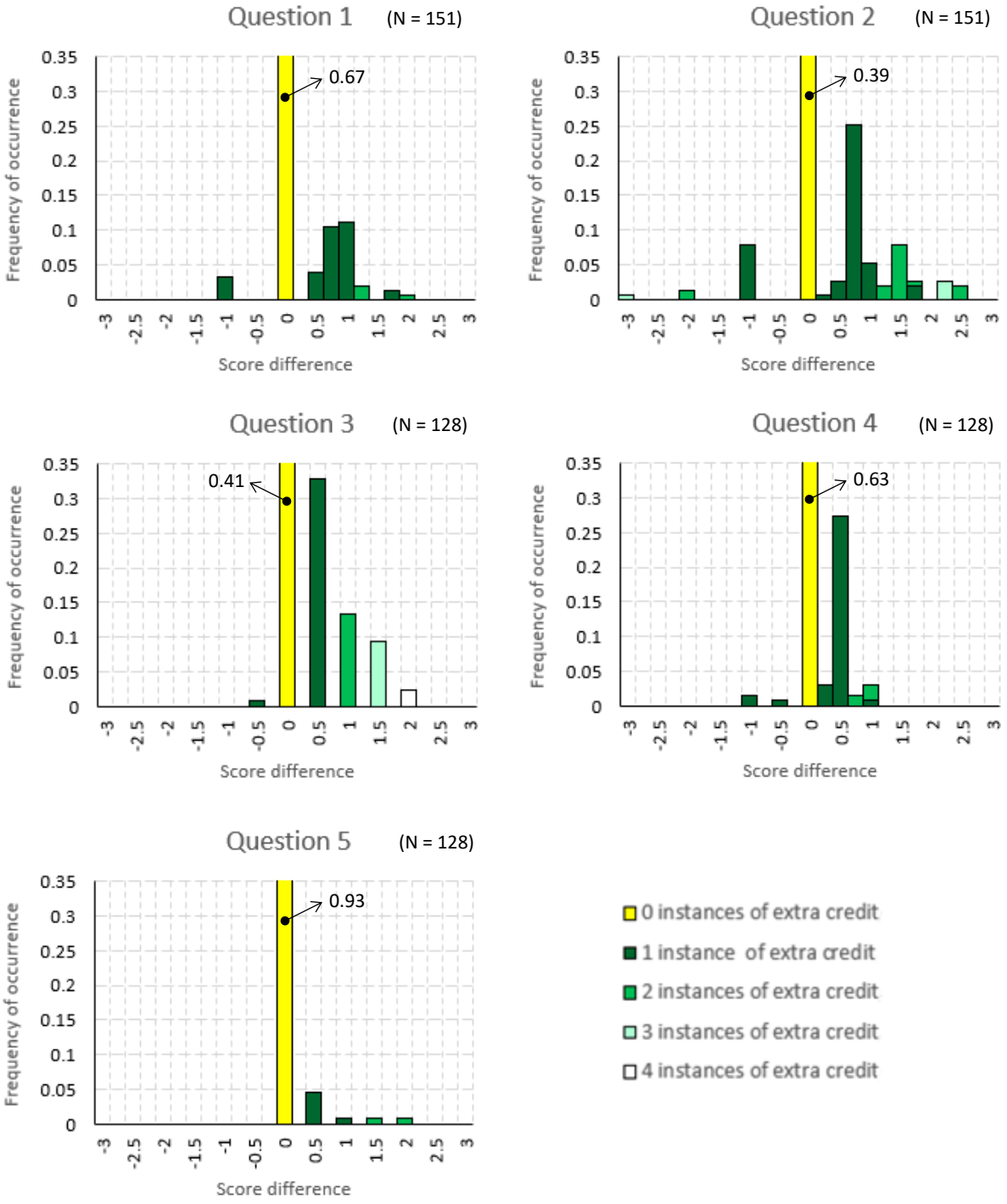



**Figure 4.** Analysis of score difference between comprehensive and correctness grading (building-block problem design).

If there is a difference, it is typically 1 point or less (questions were between 5 and 7 points, see Table 3). In those cases, students often received points for setting up a problem but then not carrying through the solution or making basic arithmetic errors (even though questions were designed to have simple math). When larger discrepancies are observed, this is mostly due to independent partial credit on several subparts of the problem (i.e., bars corresponding to more than 1 instance of partial credit). This is, for example, evident for Question 3, where some students made up to 4 mistakes that were eligible for partial credit, amounting to 2 points total. It may be hard to eliminate/reduce these situations by design.

On the other hand, for Questions 1 and 2, there are scenarios where a single partial credit was worth 1.75 points. This occurred when a mistake for one subpart propagated to another one: students were not penalized for this in comprehensive grading (i.e., a subpart was considered worthy of full credit even if the answer was wrong, when it used wrong input values from a previous subpart in the correct way). In essence, with correctness grading, students were sometimes penalized twice. The following specific cases occurred:

- Question 1: An error in solving the basics of a circuit would affect two subparts.
- Question 2: An attempt at decoupling the two subparts had been made. One subpart involved calculating an integral to find the voltage across a capacitor. The second subpart had the same integral but included an initial condition. Full credit was given for the second part if the difference with the first part was a certain value, hereby accounting for them correctly using the initial condition. However, a small error in the integral equation would affect both solutions and thus still propagated.

In both cases, more care needed to be taken to fully decouple the subparts to avoid this problem.

Total discrepancies are largest for Questions 2 and 3. For Question 2, it was largely due to the aforementioned issue, as well as the problem being challenging overall in terms of mathematics. Question 3 involved complex number calculations, which gave some students problems, even though they were kept basic. This issue mainly affected the weaker students, who generally struggled with math preparation. More generally, problems that involve more advanced skills that are not explicitly learning outcomes of the course, are harder to design for.

Instances of negative partial credit, where students received points with correctness grading but not with comprehensive grading, were all attributable to them not including any work or accidentally arriving at the correct answer via an incorrect approach. These instances were rare. However, in Question 2, for some students, this occurred independently for more than one subpart.

Overall, we see that credit lost due to correctness grading for a question was at most 2.5 points. The average score difference was 4.58% per question, with a standard deviation of 9.24%. For our final exam out 31 points, a student would on average have scored 1.4 points less if correctness rather than comprehensive grading were used, i.e., if no partial credit was awarded

for the worked solution. Assuming independence between scores on five questions, we calculated that the likelihood of seeing an additional score impact of 10% or more is below 1.25%.

***Problem Design Comparison***

We also wanted to look at the exam question design itself, i.e., evaluate whether there was a significant difference when a question was presented in a holistic way versus when it was made up of subparts corresponding to different building blocks. The conversion of a holistic question design into a building-block design is a key step in our proposed approach.

Figure 5 on the next page shows for each question the cumulative distribution function of scores achieved, under the different designs and grading strategies. The key question is the impact of switching the question design itself to a building-block one. This can be answered by comparing the holistic design (blue curve) and the building-block design with comprehensive grading (red curve). While there are discrepancies, it is important to note that we are comparing distinct variations of a question here, as each was given to one of the two sections. For academic integrity reasons, these questions could not be identical. The starting point of the methodology is a holistic version of the question, and even if we had offered this version rather than the derived building-block one, it is expected that scores would not match perfectly between sections. Even though we attempted to test the exact same concepts with a similar level of difficulty, it is likely that the two versions were not completely equivalent. Importantly, it is difficult to design two distinct exam questions that test the same knowledge and that are of the exact same difficulty level. For example, often something as small as swapping the polarity of a source, which is trivial for an expert, may trip up some more novice learners, having an impact on the average score on that question.

Furthermore, for Questions 1 and 2, the building-block version was given to Section A, where for the other three questions this version was given to Section B. Taking this into account, one notices that Section B consistently outperforms Section A. This corresponds to how these students did on other assessments in the course as well, with Section B outperforming Section A. A possible explanation is that the lecture slot for one section coincided with another class that a lot of students had to take (specifically those slightly further along in their academic track). As such, differences between the two curves in these figures may also be partly due to intrinsic performance differences between the two class sections.

We also asked the graders, who had worked for this course for prior terms, to comment on the building-block designs of the different questions. All felt that the questions were of a difficulty level similar to the holistic versions, as well as exams of prior terms, and that the exam was as well-suited to fairly assess student knowledge.

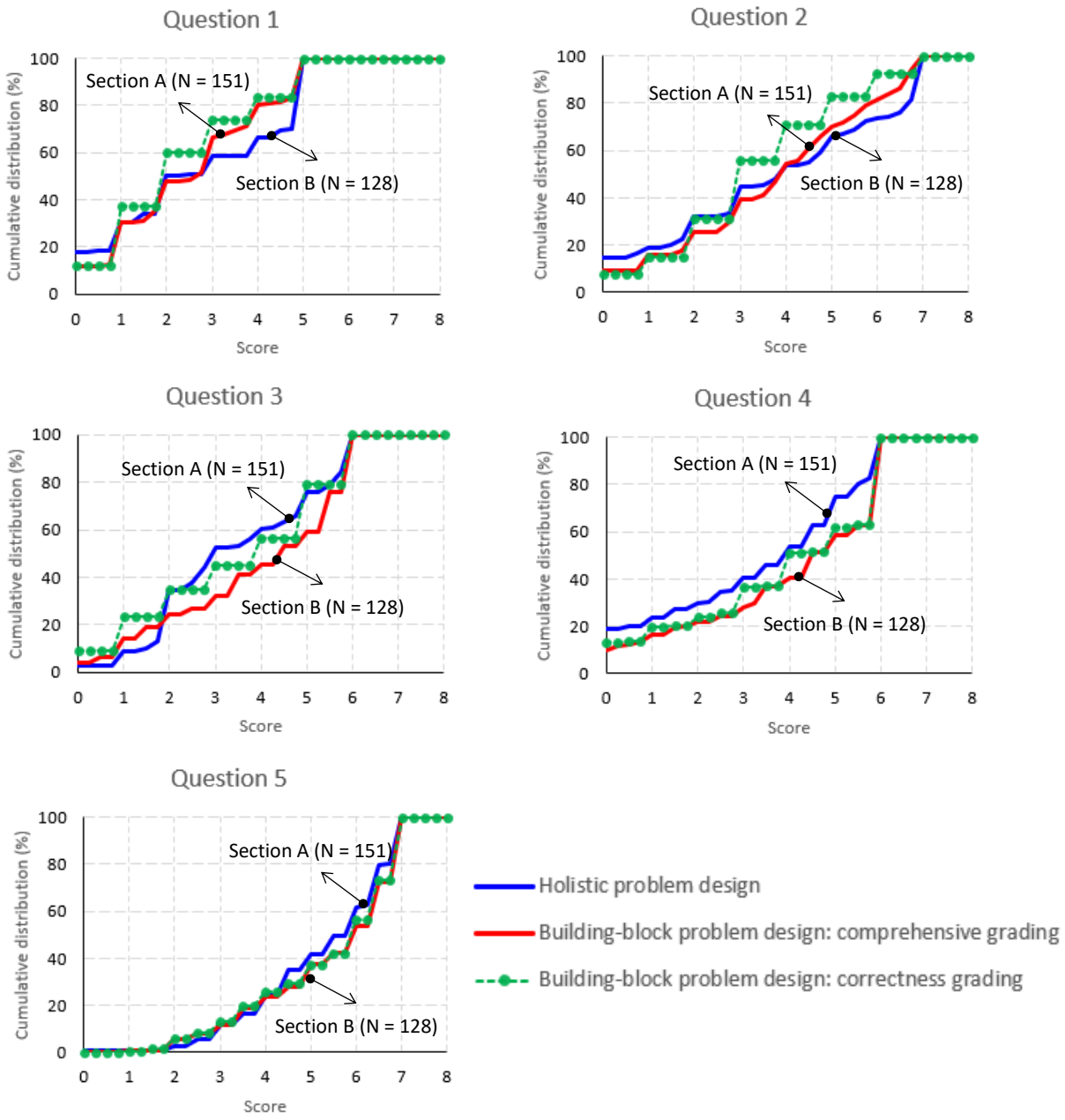



**Figure 5.** Cumulative score distribution comparing the different problem design and grading strategies.

Figure 5 also reiterates some of the observations we made earlier about the comprehensive versus correctness grading. In comparing the green and red curves, we see that they map fairly closely for most questions. As was also evident in Figures 3 and 4, and again here in Figure 5, the impact of the grading scheme is the greatest for Questions 2 and 3. Moreover, we notice that

the discrepancy because of grading scheme is often similar in extent to that due to problem design. As argued earlier, the latter may be mostly due to factors such as exam difficulty, which are almost impossible to correct for anyway, suggesting that these levels of discrepancy may be acceptable.

***Grading Effort Comparison***

Finally, we wanted to look at the grading effort involved, as the goal of the proposed methodology was to reduce this effort. Table 4 shows a comparison of the grading time for the different exam designs and grading approaches. Exams were scanned into Gradescope. Each question, whether in the holistic design or the building-block design, was assigned to one grader or to two graders working together, to grade based on worked solution. Next, correctness grading was applied to the five building-block problems, leveraging the Gradescope AI, as explained earlier. This was done by a single grader. It took them 3h 32min to process all five questions. This amounted to 2.98 seconds per subpart per student. Or for our 31-point exam, consisting of 31 subparts, it would have taken about 92 seconds per student. Note that this is an estimate based on the average, as no single student received building-block problems for each of their questions (they were distributed between the two sections).

**Table 4.** Grading time comparison.

| | Holistic problem design | | | Building-block problem design | | | |
|---|---|---|---|---|---|---|---|
| | | | | Comprehensive grading | | | Correctness grading |
| | Total grading time | Number of students | Grading time per student | Total grading time | Number of students | Grading time per student | Grading time per student |
| Question 1 | 3h 58min | 128 | 112s | 5h 47min | 151 | 138s | 15s |
| Question 2 | 5h 13min | 128 | 146s | 10h 30min | 151 | 250s | 21s |
| Question 3 | 11h 26min | 151 | 273s | 5h 15min | 128 | 148s | 18s |
| Question 4 | 6h 20min | 151 | 151s | 4h 32min | 128 | 128s | 18s |
| Question 5 | 4h 20min | 151 | 103s | 3h 47min | 128 | 106s | 21s |
| Average | | | 157s | | | 154s | 19s |

In looking at Table 3, one notices that grading times per question varied significantly. This was due to question complexity as well as grader experience. In addition, for some questions, the holistic design made keeping track of students work harder than in the building-block design. For other questions, the building-block design was longer due to it consisting of several independent parts. Overall, the average grading time ended up being very similar.

Correctness grading was much faster, representing a roughly **8-fold improvement in grading time**. The main time sink for this grading approach was that the Gradescope AI sometimes had

issues deciphering the student's handwriting. In those cases, the grader had to manually assign the results to the appropriate category. This was particularly noticeable for problems with complex numbers, where the flexibility of cartesian versus polar coordinates added additional overhead. This was the case in general when it was not specified what format final answers needed to be in; for example, "3/2" versus "1.5". All of this resulted in some manual work that had to be done by the grader to assign answers to the correct grading group. With more restrictions on answer format, we believe grading time could therefore be reduced further.

## Conclusions

In this paper, we explored a methodology to create written summative assessments that can be graded based on correctness, while still remaining representative of students' mastery of the course topics. The methodology starts from the normal grading rubric and uses it as a guide to extract the core conceptual ideas that make up the problem, which are then translated into standalone parts of the reworked question. Each of these parts is then graded based on the final answer only, while allowing for grading granularity beyond more correct/incorrect. This approach was validated for the final exam in a large undergraduate electrical engineering class, with 279 students split over two sections.

To evaluate whether the resulting exam design remained representative, we looked at the impact of the two steps in our approach: question design and grading format. Directly comparing problem designs was challenging, as the two versions were given to different student groups. However, by looking at the trends, we believe that the building-block approach can provide a valid and representative assessment, given proper care is taken in its design. Feedback from the grading team echoed this as well. When investigating the impact of only considering the final answer, in our experiment, students would lose some partial credit, on average around 5%. This could be partially compensated for by shifting the grading scale accordingly. An important component is to design problems that are easy in a numerical sense. This is an approach we had already adopted before this study: all questions were set up with easy numbers and focused on testing application of concepts rather than subjecting students to lengthy calculations. The methodology we propose here further reemphasizes this philosophy, and forces one to think about effective ways to evaluate concepts rather than the ability to do rote calculations. Also, it is important to acknowledge that ultimately, whether following this approach or the traditional holistic one, creating good questions depends on the skill and dedication of the instructor.

In addition, our methodology assumes that problems can be broken down into constituent concepts. A potential drawback is that it appears to preclude testing more high-level reasoning of *how* to approach a problem. However, even in holistic problem design, this is difficult to test – if a student makes a mistake on how to approach a complex problem, it is challenging to award partial credit in that case as well, which is problematic if the problem constitutes a significant

portion of the grade. One might argue that in this case, it is best to find the shortest possible problem to test whether students can select the correct approach. This is in fact in line with the proposed methodology, which advocates finding a problem formulation that focuses on a single concept -- in this case, the selection of the correct approach. Nevertheless, we acknowledge that this may be challenging in practice.

It is also important to note that holistic grading based on the full worked solution is not perfect itself, as there are effects of grader variation and grader error [24]. Automated grading, which is feasible when only considering final answers, can mitigate most of these issues. Importantly, it also speeds up the correctness grading process. For our implementation, we leveraged a grading AI to automatically group the open-response final answers. An alternative approach is to use multiple-choice options instead. We did not pursue this approach because of the study design, where we also needed to compare to the comprehensive grading and did not want to bias students' work by providing multiple choice answers. Additionally, multiple-choice would expose us to the problems described earlier, specifically related to student guessing and anticipatory learning (i.e., students who expect multiple-choice not to test deep knowledge may not pursue this level of knowledge) [3]. As such, we believe that our approach has more value as a constructed-response exam, where grading is based on the open-ended final response.
It can be especially useful for large classes, where grading effort is significant. Grading time that is thus saved could be reinvested into the course by offering more tutoring hours, for example. We feel that the educational benefit of helping students makes up for the loss in grading accuracy. Furthermore, as proposed in the work of Veale and Craig [24], students could be trained on how to check their answers. This would not only provide a way to mitigate some of the issues regarding losing points due to careless errors, but it would also provide an opportunity to teach and reinforce the notion of sanity checks as valuable engineering skills.

In addition, one could also address this accuracy loss by instituting a regrade policy that allows students to get credit for their worked solution. For example, one could share a rubric and allow students to submit regrade requests if the discrepancy is larger than a certain amount. In effect, this would alter comprehensive grading to become "on demand". It would allow trading off follow-up grading effort with accuracy. In addition, it would encourage students to revisit their own work, which may have learning benefits in itself. While we have not explored this option yet, it is part of future work.

## Acknowledgements

We would like to think our graduate and undergraduate Instructional Assistants, Samuel Woo, Nitya Agarwal, Joshua Orozco, Brandon Cramer, Joshua Hayes, Eddie Lu, Silvia Liu, Aaryan Tiwary, Jean Calicdan, Yizhang Liu, Matthew Alfaro, Conner Hsu, Edwin Oliveros, Jacob Lopez, and Jonathan Koby Cayaban, for helping grade the exams and providing their insights.

## References


[1] R. Pamphlett and D. Farnill. "Effect of anxiety on performance in multiple choice examination." Medical education 29.4 (1995): 297-302.

[2] P. J. Stavroulakis et al. "Comparison of Electronic Examinations using Adaptive Multiple-choice Questions and Constructed-response Questions." CSEDU (1). 2020.

[3] M. Simkin and W. L. Kuechler. "Multiple-choice tests and student understanding: What is the connection?." Decision Sciences Journal of Innovative Education 3.1 (2005): 73-98.

[4] M. Paxton, "A linguistic perspective on multiple choice questioning." Assessment & Evaluation in Higher Education 25.2 (2000): 109-119.

[5] G. R. Hancock "Cognitive complexity and the comparability of multiple-choice and constructed-response test formats." The Journal of experimental education 62.2 (1994): 143-157.

[6] P. Costa, P. Oliveira, and M. E. Ferrão. "Statistical issues on multiple choice tests in engineering assessment." Proceedings of the 37th Sefi Conference, Rotterdam, Delft University of Technology. 2009.

[7] D. Triantis et al. "Comparing Multiple-Choice and Constructed Response Questions Applied to Engineering Courses." Computer Supported Education: 6th International Conference, CSEDU 2014, Barcelona, Spain, April 1-3, 2014, Revised Selected Papers 6. Springer International Publishing, 2015.

[8] P. Photopoulos et al. "Preference for Multiple Choice and Constructed Response Exams for Engineering Students with and without Learning Difficulties." CSEDU (1). 2021.

[9] R. M. Kaipa, "Multiple choice questions and essay questions in curriculum." Journal of Applied Research in Higher Education 13.1 (2021): 16-32.

[10] C. A. Melovitz Vasan et al. "Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course." Anatomical sciences education 11.3 (2018): 254-261.

[11] W. L. Kuechler and M. G. Simkin. "Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test." Decision Sciences Journal of Innovative Education 8.1 (2010): 55-73.

[12] R. Lukhele, D. Thissen, and H. Wainer. "On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests." Journal of Educational Measurement 31.3 (1994): 234-250.

[13] B. Bridgeman, "A comparison of quantitative questions in open-ended and multiple-choice formats." Journal of Educational Measurement 29.3 (1992): 253-271.

[14] E. Ventouras et al. "Comparison of oral examination and electronic examination using paired multiple-choice questions." Computers & Education 56.3 (2011): 616-624.

[15] D. A. Bradbard, D. F. Parker, and G. L. Stone. "An Alternate Multiple-Choice Scoring Procedure in a Macroeconomics Course." Decision Sciences Journal of Innovative Education 2.1 (2004): 11-26.

[16] L. W. Anderson and D. R. Krathwohl. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman, 2021.
[17] S. A. Atwood and A. Singh. "Improved pedagogy enabled by assessment using gradescope." 2018 ASEE Annual Conference & Exposition. 2018.
[18] Electrical Technology, “Active, Reactive, Apparent and Complex Power,” https://www.electricaltechnology.org/2013/07/active-reactive-apparent-and-complex.html
[19] M. Yen. S. Karayev, and E. Wang. "Analysis of grading times of short answer questions." Proceedings of the Seventh ACM Conference on Learning@ Scale. 2020.
[20] J. L. Falconer and J. deGrazia. "Grading Exams and Homework More Efficiently and Effectively." Chemical Engineering Education 53.2 (2019): 100-100.
[21] A. Singh et al. "Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work." Proceedings of the fourth (2017) acm conference on learning@ scale. 2017.
[22] Y. Zhang, R. Shah, and M. Chi, "Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading." International Educational Data Mining Society (2016).
[23] J. Sandland and P. Rodenbough. "Strategies for Assessment in Materials Science and Engineering MOOCs: Short-Answer Grading Best Practices." Open Education Global Conference. 2018.
[24] A. J. Veale and T. S. Craig, "Design principles for final answer assessment in linear algebra: implications for digital testing." Teaching Mathematics and its Applications: An International Journal of the IMA 41.4 (2022): 280-291.
[25] C. Sangwin, "Developing and evaluating an online linear algebra examination for university mathematics." Eleventh Congress of the European Society for Research in Mathematics Education. No. 15. Freudenthal Group; Freudenthal Institute; ERME, 2019.
[26] G. Wiggins, "The case for authentic assessment." Practical assessment, research, and evaluation 2.1 (1990): 2.