International Journal of Engineering Education Vol. 37, No. 5, pp. 1441–1453, 2021
Printed in Great Britain

0949-149X/91 $3.00+0.00
© 2021 TEMPUS Publications.

Constructive Alignment Integrated Rating (CAIR) in the Assessment of Engineering Problem-Solving*

BAHAR MEMARIAN and SUSAN MCCAHAN
Dept. of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, Canada.
E-mail: bahar.memarian@mail.utoronto.ca, susan.mccahan@utoronto.ca

There is a need to characterize engineering problem-solving errors on student solutions to provide more formative information in a manner that aligns with the course and program learning outcomes. In this study a new formative feedback framework schema, designed to improve the apparency in feedback delivery, is proposed and tested. The design of the schema is inspired by the concept of work domain analysis from industrial engineering. Unlike conventional marking which uses point deductions to imply error severity, the new instrument is centered on feedback that characterizes the nature of each error. The schema was tested with teaching assistants (assessors) in a three-part semi-randomized evaluation trial. A group of engineering assessors (n = 33) evaluated problem solutions using the new schema and conventional grading. The new system significantly enhanced the proportion of descriptive feedback assessors provided across solutions ($p < 0.001$, effect size = 0.46). The new system also significantly decreased the speed of feedback delivery as compared to conventional grading ($p < 0.001$, effect size = 0.60). Conventional grading, however, scored significantly higher on the SUS usability score ($p < 0.001$, effect size = 0.73). Within the scope of the electrical engineering problem solutions tested, the new schema significantly improved the quality of formative assessment assessors provided relative to conventional marking. However, the usability of the instrument still needs improvements. Future work could include creating a digital interface based on the proposed framework to improve the evaluation experience for the assessors.

Keywords: constructive alignment; formative feedback; assessment; problem-solving; work domain analysis

1. Introduction

Closed-ended problems that rely on science and math concepts are a staple of engineering curricula. The need for problem-solving competency development is acknowledged in engineering education [1]. In this domain of engineering competency, conventional forms of assessment persist [2]. Conventional forms of marking such as numerical grading schemes, while efficient, offer little formative or transferable insight for the student on the quality of their problem-solving [3–6]. To address this issue, we are proposing a new type of schema for the assessment of engineering problem-solving skills specifically to enhance formative feedback quality. In this work, we test this approach by examining the feedback generated by assessors using the new schema.

A key goal of formative feedback delivery is to achieve constructive alignment in the assessment cycle [7]. That is, the feedback provided by the assessor(s) (e.g., teaching assistants) should align with the evaluation criteria set by the instructor based on learning outcomes for the course. To strengthen this alignment, each piece of feedback the assessors provide should be clearly linked to a specific criterion. In our work, the feedback given by people marking student solutions using the new schema needs to speak to universal dimensions of engineering problem-solving skills and meet the needs expressed in the literature.

1.1 Literature Review

There are two important bodies of literature that pertain to this work: research on problem-solving as a critical competency; and research on assessment. Effective problem-solving ability is essential for competent engineers [8]. While the context and complexity of problems vary, the need for information processing and goal-oriented behavior is common [9]. Both problem-solving specific to a particular content area as well as general problem-solving skills are needed for mastery [10]. Various works in literature have attempted to devise models and strategies related to problem-solving. Founding models of problem-solving generally utilize step-by-step elaboration as an approach to improve the quality of solutions. Building upon this work, models such as Polya [11], General Problem Solver [12], IDEAL problem solver [13], and Woods process [14] emphasize the problem solver's step-by-step communication of the solution. Most of the problem-solving models proposed are general in the sense that they decompose the process of problem-solving into actionable phases, including but not limited to, understanding the problem, planning a solution, carrying out problem-solving, and reflecting on the solution [11, 14–19]. These types of

* Accepted 2 June 2021.

frameworks are used to elicit well-structured solutions from the student, which is an important aspect of developing problem-solving. It should be noted, however, that these problem-solving models were not designed explicitly as assessment frameworks for formative feedback delivery.

The second area of research literature related to this work is assessment. It is well known that assessment can serve not only for summative evaluation purposes, but also as a learning activity through the way a problem is set, the way the solution is structured, and alignment with other aspects of the course [20]. Assessment thus becomes not only an evaluation of current ability but rather a continuous and aligned process that complements the educational experience by correcting misconceptions and facilitating learning. This *Assessment for Learning* approach benefits the student and provides feedback to the instructor. Biggs' constructive alignment framework [7] and Black and Wiliam's theoretical framework of formative assessment [21] are established conceptual frameworks in this domain. Aligned Assessment, Formative Assessment, Learning Outcomes Assessment, and Competency-Based Assessment are terms commonly associated with an *Assessment for Learning* orientation. These conceptual frameworks, and the associated concepts, maintain that assessment needs to be aligned with the real-world demands of a discipline rather than just its historical and theoretical underpinnings. For students, problem-solving activities provided over the course of the program become their practice field in preparation for the profession [22]. General principles for effective assessment include: maintaining consistent and appropriate formative communication between assessors and students [23], communicating in a manner that is purposeful and informative [5], and providing feedback that can help in adjusting learning gaps for both the task at hand and those in the future [24].

Much of the present research in assessment focus on validity and reliability considerations [25–27]. Common advice for improving assessment emphasizes either increasing grading scheme precision or creating a more detailed list of criteria (i.e., increased granularity), often in the form of a detailed feedback schema or rubric. In recent work for example, Grigg and Benson proposed a 54-item feedback scheme to provide feedback on well-structured engineering problems categorized by tasks, errors, strategies, and accuracy [28]. The list identified the conditions that need to be met for the solution components to be deemed correct. This obviously would yield very detailed feedback for a student.

The literature also notes a need to distinguish between candidates who do not know the disciplinary knowledge but who can think critically, versus candidates who know the subject but whose problem-solving approach is based on memory [10]. Tools (e.g., rubrics or problem-solving models) may place substantial value on the individual steps in the solution, which is not necessarily an indicator of overall fundamental understanding. Hull, for example, presented a case study of two students and showed that communication is not the main predictor of expertise, but rather the attainment of problem-solving skills [29]. The authors suggested that assessment tools should utilize criteria that can discriminate between performance levels based on problem-solving skills.

There is also a body of work on the differences between novice and expert problem solvers. In physics, Chi qualitatively observed several differences between these types of problem solvers [30]. For example, novices seem to skip the step of qualitatively addressing a problem before moving forward with writing factual relationships and carrying out calculations. In contrast, experts tend to write "meta-statements" and comment on the state of the problem in the solution [30, 31]. In addition, experts utilize a forward solving strategy, while novices seem to move backward. Experts start with the variables provided by the problem and identification of the nature of the problem (e.g., conservation of energy) to deduce appropriate equations to use to solve the problem. Novices, however, tend to take a "trial-and-error" approach. A new formative feedback instrument could look for these characteristics to indicate expertise in problem-solving while also supporting the development of expert problem-solving skills such as these.

1.2 Gap and Focus of this Work

Engineering problem solving has been extensively studied [18, 19] and assessment has been identified as the weakest link in learning to problem solve [32]. A long history of research in grading and feedback supports the need for improved assessment practices [2]. Most significantly, there is an identified need for: systematic research in the field [33], design of assessment instruments that enable a consistent degree of formative feedback [34], and development of assessment guides for assessors which effectively communicate pedagogical expectations for formative feedback [6]. While there have been attempts to create rubrics and other types of assessment tools that follow problem-solving steps, there is an opportunity to develop alternative tools that constructively align feedback on engineering problem-solving skills and particularly make feedback more "apparent" and consistent.

The concept of "apparency" comes from human

factors literature in industrial engineering [35]. It is used to describe systems or interfaces where the logic underlying the system is made apparent to the user by the design of the system. In our work, the concept of apparency is applied to feedback. Is the nature of the errors in a problem solution apparent from the feedback the assessor provides? We also examine feedback speed, whether the schema changes the speed of feedback delivery. By speed, we mean the number of feedback pieces provided per evaluation time spent for each solution. The design of the schema thus focuses on supporting the work done by assessors to mark papers in a way that enhances apparency and constructive alignment.

2. Theoretical Perspective/Conceptual Framework

To reimagine an assessment tool for supporting formative feedback on engineering problem-solving tasks, we started with existing theoretical frameworks to inform the design. Assessment is essentially an interaction that supports constructed understanding through social negotiation: i.e., we are taking a social constructivist perspective for the design and use of assessment. This may be obvious when we consider ideal feedback cycles, such as the cycle proposed by Hewson and Little in medical education, but it is also a useful approach in engineering education contexts [36]. Consider how each step in the assessment process can be viewed as a social negotiation intended to support meaning-making:

(a) Assessment of student learning begins with an elicitation, which is the assignment of a task or activity. The elicitation itself carries meaning as to what the instructor values.
(b) The student performs the task and submits the result to the instructor. This is a form of communication back to the instructor. It is, in essence, a question "do I understand correctly?" A well-developed elicitation and response should demonstrate the student's abilities with respect to identified learning outcomes.
(c) The student's work is graded, and feedback is provided. The feedback should support the development of meaning for the student, regardless of the solution path they have chosen.
(d) The student reflects on the feedback. This step is often identified as weak or missing in the cycle but is critical for the social negotiation of concepts. This step can be improved by providing feedback that is directly linked to intended outcomes.

Through the assessment cycle the student, and assessor(s) are socially negotiating several ideas: what concepts are important in this field of study, the meaning of those concepts, how they are correctly applied, what kinds of problem-solving are valued, what value is put on the solving process versus the right answer, and so on. In addition, in an assessment that demonstrates constructive alignment with a course and program of study [7], the feedback is aligned with broader curricular goals. The assessment cycle itself can actually be viewed as a mediating system with dynamic rules through which the social negotiation takes place.

Marking schemes that use rubrics or line-by-line identification of errors attempt to create a precise, objective form of feedback to communicate to the student areas of misunderstanding on each solution. However, by focusing on individual steps of the solution, these forms of marking have two key weaknesses; they do not easily account for alternative solution pathways, and they may lack transferable information related to problem-solving competency. This can result in feedback that is very specific to the particular problem that was assigned (e.g., specific to a circuit analysis) rather than communicating information about the student's transferable problem-solving skills. Any new tool for formative assessment should support an effective social negotiation cycle and particularly provide actionable, transferable information to the student about their problem-solving skills.

2.1 Tool Development

Inspired to achieve constructive alignment in the assessment of engineering problem-solving skills, we developed a new schema: the Constructive Alignment Integrated Rating (CAIR) system. To construct CAIR, we used a standard engineering design process informed by the field of work domain analysis. Work domain analysis investigates the purpose and structure of a work domain in which users operate [37]. The assessment material, guides, and so on (e.g., problem sets, marking scheme, etc.) collectively make up the work domain in this case, and the assessors (i.e., Teaching Assistants) are the users. The literature on assessment suggests that descriptive (i.e., elaborative) feedback is more valuable than basic corrective feedback, which often takes the form of check-marks or cross-marks with no explanation [34]. However, elaborative feedback in the form of sentences or phrases can be time-intensive. Analytic rubrics, although rarely used in problem-solving assessments, offer one example of a design that tries to address this challenge. Our design attempts to combine the efficiency of a rubric with the results of a work domain analysis to create scheme designed for transferable formative feedback delivery.

To utilize criteria for evaluation and feedback delivery that are common irrespective of the particular problem posed (e.g., a circuit problem or a statics problem) we identify aspects of engineering problem-solving that are common across disciplines. From the work in human factors, it has been established that expert work, like engineering problem-solving, has a hierarchical configuration. This is referred to as a "why-what-how" or means-ends process [38]. In the work by Rasmussen and Vicente, this is referred to as levels of abstraction [38], [39]. Levels can represent the same engineering problem-solving activity, but from different unique perspectives; essentially different abstractions of the activity that all load onto the same construct (i.e., problem-solving ability). One level is not necessarily more important or more valuable than another. Together they create a deeper and more complete view of problem-solving. Levels of abstraction can help us to differentiate between student solutions that have (or lack) critical thinking, a grasp of engineering theory, and a grasp of mathematical accuracy. Attempting to address needs reported in the literature concerning critical thinkers versus problem solvers who chiefly rely on memory [10], we discriminate between the following levels of error abstraction:

1. Goal: The masterplan, goal, and strategy of a problem solution. This level provides the most abstract view of problem-solving. Students need to identify the overall goal of the problem, and the resources available (i.e., a strategy as opposed to specific steps). This level can be understood as the purpose-related function (key problem-solving components) students execute to move from an ambiguous problem to a solved solution. In a problem solution, it is manifested by identification of the unknowns and knowns as they appear throughout the solution.
2. Theory: Engineering theories and principles that govern the problem. This level provides a theoretical view of the underlying engineering conceptual model. This level focuses on the engineering related principles that govern the problem and can be understood in terms of engineering-related formulas and associated disciplinary standards (i.e., constraints and assumptions).
3. Calculation: Computational techniques and calculation work. This level is analogous to the mechanical process of solving a problem. It is an integral component of traditional grading practice; the mathematical and numerical view of problem-solving which focuses entirely on computational work accuracy. This third level can thus be assessed through numbers, mathematical procedures, and the condition of results evident in the solution.

In addition, at each level, we can identify a decomposition that allows the assessor to look for markers that indicate expert problem solving throughout the solution, not just at each step. There are numerous ways to characterize a system into levels of abstraction and decomposition. Work domain analysis methodology suggests selecting the levels based on the underlying needs of the work domain. In this research, we approach the feedback process as the work domain, with the need for high-quality transferable feedback being the primary goal. In studies on expertise, an emphasis is placed on the attributes that distinguish between novice and expert problem solvers [10, 30, 31, 40–44]. Most commonly, expert problem solvers demonstrate greater proficiency in identifying conceptual knowledge (i.e., knowns and unknowns involved), theoretical underpinnings of problems, and the most relevant strategy for solving a problem. Novices, on the other hand, demonstrate an irrelevant, incomplete, or inaccurate grasp of problem-solving and the nature of the problem they are solving. These findings would suggest that novice problem solvers are more likely to make "deep errors"; that is, errors that suggest a fundamental misunderstanding of the problem and solution space. Expert problem solvers, conversely, are more likely to make no errors, or only "surface errors"; that is, errors that are superficial such as minor calculation errors.

Building on this, the levels of decomposition attempt to separate a lack of precision (i.e., a bit of carelessness) in a problem solution from fundamental flaws in conceptual understanding manifested in the problem solution. The proposed schema is intended to distinguish a student's depth of ability from an artifact (solution) that is both deeply rooted in context, that is the specifics of the particular problem posed, and is often devoid of commentary, that is the student will rarely explain their thinking alongside their work. This is inherently imprecise, so we chose three levels of abstraction and two levels of decomposition rather than a more granular system. A previous analysis of errors made by students on problems of this kind demonstrated that the proposed schema could capture all of the types of errors commonly found on problem solutions [45]. We realize that the criteria could be more detailed. However, our proposed scheme is intended to identify the nature of problem-solving errors in engineering using the smallest number of error types for simplicity. Any further granularity would suggest that the instrument is more precise than it really is.

2.2 Walk Through of CAIR

We are calling the instrument CAIR (constructive alignment integrated rating) because it is intended to achieve objectives of constructive alignment and formative assessment. Putting the dimensions of the abstraction hierarchy together with the decomposition levels we propose a two-dimensional matrix form (see Fig. 1.). In traditional grading, every error in a solution may receive a cross-mark and usually a point deduction and/or a piece of elaborative feedback (e.g., "wrong equation") [45]. CAIR utilizes a qualitative error classification system instead. The work of assessors entails identifying the errors and their type (i.e., surface or deep, and goal, theory, or calculation). The performance rating is then determined by the error type(s) detected.

The rows represent the abstraction levels of goal, theory, and calculation and the columns decompose each abstraction level into deep and surface type errors. To differentiate between CAIR items efficiently, alphabetic symbols are used to represent each cell (see Fig. 1). The alphabetic symbol represents the first letter of that error category; U is for unknowns, and so on. "Unknowns", for example, does not simply mean did the student find the unknowns that are elicited in the problem statement (i.e., the goal of the problem-solution), but rather refers to the overall strategy used by the student to reach the goal.

At the goal level of abstraction, assessors look for the overall flow of problem-solving and ways in which known variables are used and unknown variables are derived. This level is the metacognitive process view of a solution and examines whether the trajectory of problem-solving presented is correct. At this level, deep errors are related to incorrect identification of unknown variables (goals of the problem): did the student devise a trajectory toward the correct goal? Or are there errors, (e.g., they identified the wrong goal or solved for the wrong thing)? Surface errors are related to incorrect identification of known variables (typically provided in the problem). At the theory level of abstraction, assessors look for errors in the use of engineering models and principles that govern the problem and associated disciplinary standards. At this level, deep errors are related to incorrect grasp of engineering models, while surface errors are related to incorrect identification of disciplinary standards and constraints within a given engineering framework. For example, if a student draws an incorrect free body diagram, this demonstrates a lack of understanding of the theoretical framework. While if the student finds the force of a mass by using the wrong units (e.g., mass in grams rather than kg), it would be a lack of understanding of disciplinary standards. At the calculation level of abstraction, assessors examine the quality of the computations in the solution. The deep aspect of computation is related to the appropriateness of the approach adopted (e.g., integration when there should be differentiation or incorrect integration) while surface errors would be purely arithmetic.

The novelty of the CAIR schema is that it can be used to explicitly identify the type and depth of each error. In traditional grading, this is often implied through the point deductions (e.g., 1 point off for an arithmetic error, and 5 off for a wrong equation),

	Deep Decomposition	Surface Decomposition
Goal Abstraction: "Problem-solving strategy/components"	**U**nknown variables *E.g., Found variables problem had not been asked for.*	**K**nown variables *E.g., Did not use the necessary variables problem provided.*
	Fails○ Below○ Meets○ Exceeds○	
Theory Abstraction: "Engineering Conceptual Models"	**T**heoretical model *E.g., Wrote an incorrect expression/formula for theory.*	**D**isciplinary standards within the model *E.g., Did not follow assumptions, conversions of the theoretical model.*
	Fails○ Below○ Meets○ Exceeds○	
Calculation Abstraction: "Mathematics"	**C**omputational model *E.g., the Mathematical model's rules were violated.*	**A**rithmetic work within the model *E.g., Arithmetic calculations made incorrectly.*
	Fails○ Below○ Meets○ Exceeds○	

Fig. 1. Proposed CAIR feedback framework.

but is not explicitly communicated to the student. As shown in Fig. 1, each level of abstraction contains a performance rating scale: fails expectations to exceeds expectations. This performance rating can be based on the number and type of errors (deep and surface) identified at the abstraction level using a formulaic approach: no errors should receive an "exceeds expectations" evaluation, surface errors only should be "meets expectations", etc. The performance rating at each abstraction level works to translate specific errors on the problem solution into transferable information about problem-solving competency.

This approach is less concerned with providing precise grading (i.e., putting an X next to every error), but rather potentially offers apparency about the value that an instructor is placing on various aspects of problem-solving. An example of how a student solution would be graded using conventional marking and CAIR is shown in Appendix A. The alphabetic symbols (U, K, T, D, C, and A) can be used alongside conventional grading to tag each error, communicate the type of error and support the weight of the associated point deduction. In addition, it can be used as a rubric to indicate to the student the overall quality of their work at each abstraction level: for example, are they failing, below, meeting, or exceeding expectations in their calculation quality or application of theory?

3. Methods for Testing

A marking schema that supports apparency should result in an increase in descriptive, and a decrease in corrective, feedback given by the assessors. The purpose of this research is to examine whether the proposed formative feedback framework (CAIR) supports apparency and reduces speed in feedback delivery when compared to conventional marking. In addition to apparency, a more consistent tool should maintain the volume and pace of feedback provided by the assessors.

3.1 Setting and Participants

Before conducting the study, a research protocol was approved by the Research Ethics Board (ID: 35223 and 37507). To test the performance of assessors using CAIR and conventional marking, three-part counterbalanced evaluation sessions were conducted using groups of electrical engineers. There was a total of 33 participants. The population of interest for this research study is Electrical Engineering (EE) students who qualify to be teaching assistants. The participants evaluated test items from 3 courses in the EE program. Participants were screened and only those eligible to serve as Teaching Assistants for the courses used in this study were selected to participate.

In each of the three parts of the session, each assessor was given a package of material. Each package contained one engineering problem, an associated ideal solution, and four ungraded student solutions to the problem. When grading each problem solution, the assessors were asked to record their start time and finish time. The material contained problems from actual electrical engineering exams and actual ungraded student solutions that were collected from a course. When using a conventional marking approach assessors were also provided with a grading scheme set by the instructor that consisted of notations on an ideal solution indicating point values for each step. The assessors marked 3 packages, 1 in each part of the trial, using either the CAIR feedback framework or using the conventional method for each part. The order was semi-randomized: Each assessor was randomly assigned to CAIR or conventional grading for part 1, then required to use the other method for part 2. They were again randomly assigned to one of the two tools in part 3. The data gathered in part 3 are used for confirmatory purposes.

3.2 Data Collection

After signing the consent form and prior TA experience questionnaire, training was provided by the session moderator (first author) on how to use each tool. Assessors were explicitly asked to provide elaborating feedback for every error regardless of the marking approach. Assessors were given 30 minutes to complete each part. Immediately after completion of each part, assessors were given 5 minutes to complete a survey about their experience followed by a short break. In the survey, assessors reflected on the usability of the approach and strategies they adopted. The surveys used in this study adapted items from the Standard Usability Survey (SUS) and The NASA Task Load Index (NASA-TLX) instruments for usability [46, 47]. The SUS is a 10-item survey that is intended to collect usability perceptions (e.g., ease of use, confidence in use) on a 5-point scale ranging from strongly disagree to strongly agree. The NASA-TLX is a 5-item survey intended to measure mental workload (e.g., performance, frustration level) [46, 47]. At the end of the third part the same survey was provided, and assessors were asked to reflect on their overall experience with each grading approach. All tasks were carried out by each participant individually.

Each part contained one electrical engineering (EE) problem selected from either: EE1 (a first-year circuits course), EE2 (another first-year circuits course), or EE3 (a second-year electromagnetics

course). Problems from the circuits courses (EE1 and EE2) covered concepts such as expression of induced current, mesh and nodal analysis, low/high pass filters, capacitance and inductance, phasors, and power calculations. For EE3, topics covered were dielectric materials, conduction/displacement current, incident magnetic field, and self/mutual inductances. The total marks for each problem ranged from 5 to 10 out of 100 on the exam from which the problems were drawn.

3.3 Data Analysis

The feedback given by the assessors on the student solutions was coded using a classification system [45]. Each piece of feedback was categorized as descriptive or corrective. Descriptive feedback is textual notations on a student solution. This included a note or phrase related to a marking scheme (i.e., tagging) or simply a description related to the nature of the error (i.e., elaborating). Corrective feedback is symbolic notations on a student solution, such as a check-mark (i.e., validating), cross-mark (i.e., flagging), or grade deduction (i.e., penalizing). Descriptive feedback has higher apparency than corrective feedback because it describes, or indicates, the nature of the error, and corrective feedback does not.

Utilizing this coding framework our goal was to compare CAIR to conventional marking on:

(1) Apparency
   (a) Relative descriptive ratio: the number of descriptive feedback instances divided by the total number of feedback instances on each solution.
(2) Pace of Feedback
   (a) Speed of feedback delivery is measured as the number of feedback instances divided by the time spent per solution.
   (b) Variability in the quantity of feedback provided; measured by Interquartile Range (IQR) across solutions.
(3) Usability: Self-reported ratings of the assessor experience.

In addition to having access to the ungraded student solutions in these three electrical engineering courses, we also had access to the same student solutions after they had been graded during the term by the teaching assistants in the respective courses. This allowed us to examine the summative evaluation and compare the grades that were given to the students with the grading done during our trials. We ran a validation and inter-rater consistency check on the feedback the participants provided to determine whether the framework appropriately differentiates the quality of engineering problem solvers and whether the inter-rater reliability of the framework falls within the acceptable range reported in the literature.

4. Results

The participants had a range of experience levels typical of teaching assistants at this institution: 27% indicated a high level of experience (more than two years), 20% average (one to two year), 23% low (less than a year), and 30% had no previous TA experience. Three participants did not submit the preliminary TA experience questionnaire but had a background (e.g., senior electrical engineering students or graduate students) necessary to qualify them to take part in the study.

Overall, CAIR was used to mark 208 student solutions and conventional grading was used on 188 student solutions. The difference between these numbers is because assessors were randomly assigned to either CAIR or conventional grading for part 3 of each session. Across the 208 solutions marked using the CAIR instrument, there were a total of 772 coded feedback instances resulting in approximately 4 pieces of feedback per solution on average. In contrast, conventional marking produced 932 feedback instances resulting in approximately 5 pieces of feedback per solution on average. However, although conventional grading produced more feedback in absolute terms than CAIR, the nature of the feedback was quite different. As might be expected, CAIR produced more tagging feedback because assessors used the CAIR categories to tag errors as a way of providing more information about the nature of each error. Tagging was the most common type of feedback used, with 385 out of 772 instances or 50% of the feedback in this category, followed by validating (30%) and elaborating (8.8%).

Thus, the total descriptive instances (tagging plus elaborating) using CAIR constituted approximately 59% of the feedback. In contrast, conventional marking resulted in nearly equal amounts of penalizing (32%), validating (29%), and elaborating (27%) feedback, and very little tagging (3.0%). From these data, it can be seen that CAIR produced more than double the amount of descriptive feedback per solution, on average. Feedback using conventional grading was largely penalizing or validating (i.e., corrective) without providing additional information regarding the nature of the errors. Both approaches to marking resulted in very few instances of flagging (i.e., cross-marks), possibly because assessors felt that a point deduction (i.e., penalizing) or tagging the error was sufficient to draw the student's attention to where the error occurred in the solution.

A summary of the feedback that assessors gave in

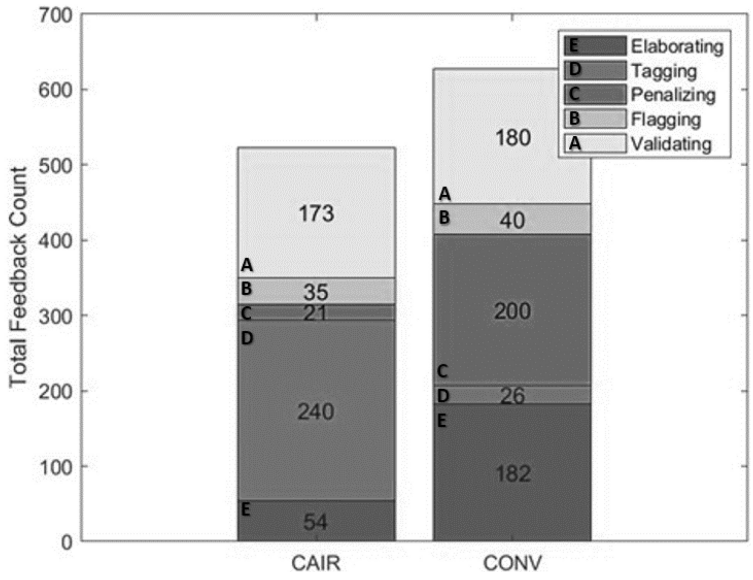


Fig. 2. Summary of total feedback quantity over parts 1 to 2 for CAIR and conventional marking (CONV).

the first two parts of each trial is shown in Fig. 2. Focusing on the first two parts of the sessions is useful because there is an equal number of solutions marked using each method which allows a visual comparison. As Fig. 2 shows, and as was noted already, conventional marking results in more feedback in general, but a lower quantity of descriptive feedback (i.e., tagging plus elaborating). It is interesting to note that both methods elicited approximately the same number of flagging and validating feedback instances. The difference in the results is largely because of a difference in assessor behavior related to tagging, elaborating, and penalizing. When using the CAIR instrument assessors seem to replace penalizing and elaborating with tagging. This may be a natural result of using a rubric like schema. However, it appears that the CAIR categories are well aligned with assessor needs, because otherwise, we might have seen the same or more elaborating feedback instances to complement the tagging.

A summary of the counts based on the quantity of feedback per solution is shown in Table 1. All solutions used in this study had at least one feedback notation, but many of the solutions quite good (i.e., received a passing grade from the course TA). As the table shows, the percentage of solutions that received no descriptive feedback using conventional grading was almost the same as when CAIR was used (27% and 24%). The major difference was in the amount of corrective feedback. Using CAIR more solutions received no corrective feedback. However, the number of solutions receiving three or more descriptive feedback notations per solution doubled when CAIR was used. The number of solutions receiving two or fewer descriptive feedback notations was almost the same between the two tools.

We calculated the percentage of feedback on each solution that is descriptive by adding elaborating and tagging notations dividing by the total number of pieces of feedback on each solution. The number of solutions where more than 50% of the feedback was descriptive are shown in Table 2. These results further support the conclusion that a larger pool of solutions (79 out of 118 that received some feedback) received more descriptive than corrective feedback per solution with CAIR. While conventional grading yielded a preponderance of corrective feedback (73 out of 119 that received feedback).

The findings from a non-parametric analysis of the data revealed that CAIR significantly enhanced the apparency of the feedback (proportion of descriptive feedback) across solutions having vary-

Table 1. Solution counts based on feedback frequency. (CONV is conventional marking)

Number of pieces of feedback:		**0**	**1**	**2**	**3**	**4**	**5 or more**	**Total # of solutions**
CAIR	Descriptive	32 (24%)	19 (14%)	29 (22%)	27 (20%)	10 (8%)	15 (11%)	132 (100%)
	Corrective	70 (53%)	16 (12%)	13 (10%)	10 (8%)	8 (6%)	15 (11%)	132 (100%)
CONV	Descriptive	35 (27%)	38 (29%)	29 (22%)	19 (14%)	7 (5%)	4 (3%)	132 (100%)
	Corrective	20 (15%)	18 (14%)	16 (12%)	22 (17%)	24 (18%)	32 (24%)	132 (100%)

Table 2. Solution counts based on relative descriptive ratio value. (CONV is Conventional marking)

Relative descriptive feedback ratio	≥50%	<50% and >0%	None (both descriptive and corrective are 0)	Total
CAIR number of solutions	79 (60%)	39 (29%)	14 (11%)	132 (100%)
CONV number of solutions	46 (35%)	73 (55%)	13 (10%)	132 (100%)

ing quality and time spent on evaluation ($p < 0.001$, effect size = 0.46). CAIR also significantly increased the amount of time spent on marking ($p < 0.001$, effect size = 0.60). This is illustrated in Fig. 3 (a) and (b). In addition to looking at the types of feedback assessors provided, we also examined the usability of each approach using surveys and the speed of marking. In terms of usability, conventional grading had a significantly higher SUS usability score ($p < 0.001$, effect size = 0.73) than CAIR, see Fig. 3 (c). Based on the NASA-TLX scores, CAIR is more taxing both mentally and physically than conventional grading. This might, in part, be due to the introduction of a new approach to marking. However, it is a finding that will need to be taken into account if we develop this schema further. Time spent marking each solution was not significantly different for the two approaches. However, CAIR significantly decreased the total marking speed (number of feedback notations provided for time spent evaluating each problem solution) across solutions as compared to conventional grading ($p < 0.001$, effect size = 0.60).

In addition to the comparative analysis of CAIR versus conventional grading, we also examined whether CAIR supports summative evaluation. This was assessed by comparing the feedback provided using CAIR with the grade that the student solution received from the actual Teaching Assistant (TA) in the course where the paper was

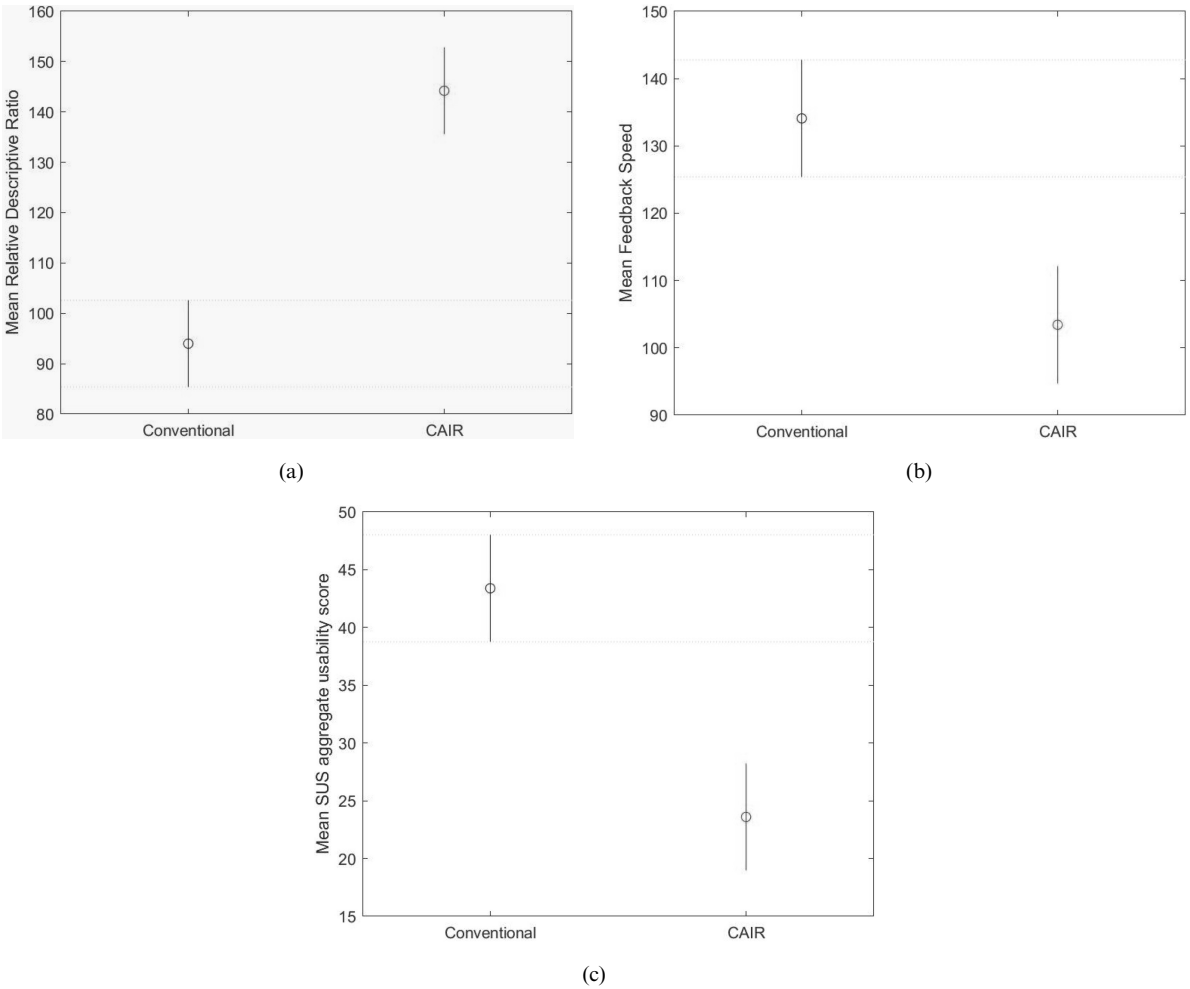


Fig. 3. Comparison tests of (a) Relative Descriptive Ratio, (b) Feedback Speed, (c) SUS aggregate usability score for CAIR versus conventional marking based on the data collected from parts 1 and 2 combined shows significant differences in the means.

Table 3. Feedback identified in fails and passing solutions.

	Fail, n = 55			**Pass, n = 135**		
	Deep	**Surface**	**Total**	**Deep**	**Surface**	**Total**
Goal	15	10	25 (22%)	11	15	26 (13%)
Theory	43	10	53 (46%)	55	31	86 (45%)
Calculation	17	20	37 (32%)	38	43	81 (42%)
Total	75 (65%)	40 (35%)	115 (100%)	104 (54%)	89 (46%)	193 (100%)

marked during the semester. Based on the literature [10], we hypothesized that:

(1) Failing solutions have more error types
(2) Failing solutions have more deep errors as compared to surface errors
(3) Failing solutions have more errors at the goal and theory levels of abstraction

The tagging (i.e., U, K, T, D, C, and A) used by the assessors during the CAIR parts of the trials was used to identify the types of errors present. We did not count the frequency of each tag on an individual solution, but rather whether a particular tag was present, or not. Using the grading done by the course TAs during the semester as a frame of reference, student solutions were divided into two groups: passing (> = 50%) and failing (< 50%). For each solution, we identified which of the six CAIR error categories was identified by the assessors during the research trials.

After removing solutions with no error tags, a total of 55 instances of failing solutions and 135 instances of passing solutions were used in the following analysis. The feedback provided via CAIR for passing and failing solutions overall (Total) and across levels of decomposition (Deep, Surface) and abstraction (goal, theory, calculation) is summarized in Table 3. The results show the alignment between the grade given by the course TAs and the tagging done via the CAIR instrument. The assessors in our testing gave failing solutions a significantly higher number of CAIR error-tag types (115 error types on 55 solutions) than passing solutions (193 error types on 135 solutions). This difference is significant ($p < 0.001$). The failing solutions were given more deep level error tags, as well as theory and goal level error tags compared to the passing solutions ($p < 0.001$). The surface and calculation levels, however, were not significantly different between failing and passing solutions ($p = 0.57$ and $p = 0.66$ respectively).

5. Discussion

A group of 33 assessors tested CAIR versus conventional grading using a large collection of electrical engineering problem-solving student solutions. The findings revealed that using CAIR significantly reduced the quantity of corrective feedback and significantly increased the quantity of descriptive feedback as compared to conventional grading. Further, the relative ratio of descriptive to total (descriptive plus corrective) feedback provided per solution increased significantly when assessors used CAIR. The results suggest that CAIR increases the apparency of the feedback provided by assessors relative to conventional marking. While the amount of time spent marking with each tool is approximately the same, assessors self-reported that conventional grading is more usable (based on SUS score) and less demanding on some aspects of mental workload (based on NASA-TLX item ratings).

Although the assessors using CAIR were given an ideal solution, but not a grading scheme, further analysis verified that the CAIR tagging convention enables differentiation between solution quality, as suggested by some researchers [30]. CAIR appropriately identified that failing solutions have a significantly higher total number of error types. Further, deep errors, which represent a lack of understanding of the fundamentals, were generally present in failing solutions. Failing solutions also had a significantly higher quantity of errors at the goal and theory levels of abstraction as compared to passing solutions. These results suggest that CAIR has content validity as an assessment instrument. Overall, the instrument tested in this work can help alleviate some of the gaps reported in formative assessment. In addition, the coding approach we used can be repurposed for systematic research in the field, specifically the type and quantity of feedback provided by assessors with different levels of experience on student solutions across different problem-solving activities [48].

The CAIR design can support the delivery of feedback that is formative, consistent, and better representative of learning outcomes [6, 34, 49] related to the engineering problem-solving process. The point of CAIR is not to produce a mark on a student's solution per se, but primarily to support meaning-making: that is, the meaning a student constructs about their problem-solving strengths and weaknesses based on the feedback. In addition, because this same tool can be used in a wide variety of engineering problem-solving courses, students

could receive cohesive feedback across their courses thus aiding in constructive alignment [7]. However, to demonstrate this would require testing the schema with students.

The use of a tool, such as CAIR, focuses the assessor on the importance of providing feedback as the primary purpose of marking. Then the subsequent process of assigning a numerical grade becomes procedural and could even be automated which, in turn, could reduce assessor workload. An instructor could assign weights to the abstraction and decomposition levels to assist in assigning a numerical grade. In addition, using CAIR or a similar instrument supports the evaluation of multiple alternative solution pathways. It eliminates the need for an instructor to create consistent grading schemes for every possible solution path.

To fully develop this approach, more testing needs to be done of the CAIR instrument. In this study, only foundational Electrical Engineering courses were used. To investigate whether the schema has a more universal application it would need to be tested in a wider range of courses and with students. It is also clear from the results that there are some usability challenges with the use of CAIR. The next iteration of this design should try to account for these issues, possibly by creating a digital form that reduces the mental load for the assessor. However, it may be that asking assessors to provide more useful feedback is inherently more taxing, but, from a formative feedback perspective, well worth the effort.

6. Conclusion

In this design-based research, we proposed a novel tool, CAIR, to enhance the quality of formative feedback on engineering problem-solving tasks. The design of CAIR was inspired by human factors and cognitive engineering concepts (i.e., Work domain Analysis, Abstraction Decomposition space). We tested the performance of assessors when using the proposed CAIR tool relative to conventional grading practice (i.e., grading scheme) on the evaluation of closed-ended electrical engineering problem-solving tasks. The results of the within measures (n = 33) counter-balanced randomized evaluation sessions reveal that CAIR significantly improves the relative ratio of descriptive feedback across solutions having varying quality and time spent on evaluation. In contrast, conventional grading had a significantly higher SUS usability score. In this work, we built on the ideas of Biggs and others in the field of Constructive Alignment, who identified that assessment and feedback are most effective when they are aligned with the intended learning outcomes for the course, and program of study. While CAIR appears to achieve the primary objectives of formative assessment and apparency, its usability still needs to be improved. A next step could be the design of a digital interface based on the findings here to improve further the evaluation experience for assessors and test the effectiveness of this feedback with students.

References

1. D. Jonassen, J. Strobel and C. B. Lee, Everyday problem solving in engineering: Lessons for engineering educators, *J. Eng. Educ.*, **95**(2), pp. 139–151, 2006.
2. S. M. Brookhart, T. R. Guskey, A. J. Bowers, I. H. McMillan, J. K. Smith, L. F. Smith, M. T. Stevens and M. E. Welsh, A century of grading research: Meaning and value in the most common educational measure, *Rev. Educ. Res.*, **86**(4), pp. 803–848, 2016.
3. R. Butler and M. Nisan, Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance, *J. Educ. Psychol.*, **78**(3), p. 210, 1986.
4. P. T. Knight, Summative assessment in higher education: practices in disarray, *Stud. High. Educ.*, **27**(3), pp. 275–286, 2002.
5. J. Dolin, P. Black, W. Harlen and A. Tiberghien, Exploring relations between formative and summative assessment, in *Transforming assessment*, Springer, Cham, pp. 53–80, 2018.
6. M. Yorke, Summative assessment: dealing with the 'measurement fallacy,' *Stud. High. Educ.*, **36**(3), pp. 251–273, 2011.
7. J. Biggs, Enhancing teaching through constructive alignment, *High. Educ.*, **32**(3), pp. 347–364, 1996.
8. S. Sheppard, A. Colby, K. Macatangay and W. Sullivan, What is engineering practice?, *Int. J. Eng. Educ.*, **22**(3), p. 429, 2007.
9. D. H. Jonassen, Toward a design theory of problem solving, *Educ. Technol. Res. Dev.*, **48**(4), pp. 63–85, 2000.
10. D. R. Woods, A. N.Hrymak, R. R. Marshall, P.E. Wood, C.M. Crowe, T. W. Hoffman, J. D. Wright, P. A. Taylor, K. A. Woodhouse and C. K. Bouchard, Developing problem solving skills: The McMaster problem solving program, *J. Eng. Educ.*, **86**(2), pp. 75–91, 1997.
11. G. Polya, *How to solve it*, New Jersey: Princeton University Press, 1957.
12. A. Newell and H. A. Simon, Human Problem Solving, in *Englewood Cliffs*, **104**(9), Englewood Cliffs, NJ: Prentice-Hall, 1972.
13. J. D. Bransford and B. S. Stein, *The IDEAL problem solver: A Guide for Improving Thinking, Learning, and Creativity*. WH Freeman, New York, NY, 1984.
14. D. R. Woods, An Evidence-Based Strategy for Problem Solving, *J. Eng. Educ.*, **89**(4), pp. 443–459, 2000.
15. G. Polya, *Mathematical discovery Vol II.: On understanding, learning, and teaching problem solving*, 1965.
16. J. E. Stice, Teaching Problem-Solving Skills, *Spectrum*, pp. 16–17, 1982.
17. A. Whimbey, Students can learn to be better problem solvers, *Educ. Leadersh.*, **37**(7), pp. 560–565, 1980.
18. D. R. Woods, T. Kourti, P. E. Wood, H. Sheardown, C. M. Crowe and J. M Dickson, Assessing Problem-Solving Skills Part 1: The Context for Assessment, *Chem. Eng. Educ.*, **35**(4), pp. 300–7, 2001.

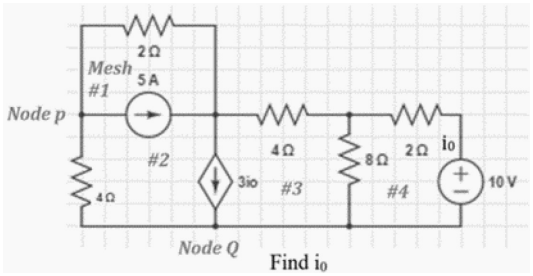
19. D. R. Woods, T. Kourti, P. E. Wood, H. Sheardown, C. M. Crowe and J. M. Dickson, Assessing Problem Solving Skills Part 2: Assessing the Process of Problem Solving, *Chem. Eng. Educ.*, **36**(1), pp. 60–67, 2002.
20. G. Gibbs, How assessment frames student learning, in *Innovative assessment in higher education* ed. Bryan, C. Clegg, K., Routledge, pp. 43–56, 2006.
21. P. Black and D. Wiliam, 'In praise of educational research': Formative assessment, *Br. Educ. Res. J.*, **29**(5), pp. 623–637, 2003.
22. G. Gibbs, *Using assessment to support student learning*, Leeds Met Press, 2010.
23. P. Orsmond, S. Merry, and K. Reiling, Biology students' utilization of tutors' formative feedback: a qualitative interview study, *Assess. Eval. High. Educ.*, **30**(4), pp. 369–386, 2005.
24. M. Ellegaard, L. Damsgaard, J. Bruun and B. F. Johannsen, Patterns in the form of formative feedback and student response, *Assess. Eval. High. Educ.*, **43**(5), pp. 727–744, 2018.
25. J. Tisi, S. Maughan and N. Burdett, A review of literature on marking reliability research. Ofqual/13/5285, 2013.
26. B. M. Moskal, J. A. Leydens and M. J. Pavelich, Validity, reliability and the assessment of engineering education, *J. Eng. Educ.*, **91**(3), pp. 351–354, 2002.
27. K. A. Douglas and Ş. Purzer, Validity: Meaning and relevancy in assessment for engineering education research, *J. Eng. Educ.*, **104**(2), pp. 108–118, 2015.
28. S. J. Grigg and L. C. Benson, A coding scheme for analysing problem-solving processes of first-year engineering students, *Eur. J. Eng. Educ.*, **39**(6), pp. 617–635, 2014.
29. M. M. Hull, E. Kuo, A. Gupta and A. Elby, Problem-solving rubrics revisited: Attending to the blending of informal conceptual and formal mathematical reasoning, *Phys. Rev. Spec. Top. Educ. Res.*, **9**(1), p. 010105, 2013.
30. M. T. Chi, R. Glaser and E. Rees, Expertise in problem solving, in *R. Sternberg (Ed.), Advances in the psychology of human intelligence*, NJ: Erlbaum, 1982.
31. E. B. Coleman and M. Shore, Problem-solving processes of high and average performers in physics, *J. Educ. Gift.*, **14**(4), pp. 366–379, 1991.
32. D. Jonassen, *Learning to solve problems: A handbook for designing problem-solving learning environments*, Routledge, 2010.
33. B. M. Olds, B. M. Moskal and R. L. Miller, Assessment in engineering education: Evolution, approaches and future collaborations, *J. Eng. Educ.*, **94**(1), pp. 13–25, 2005.
34. V. J. Shute, Focus on formative feedback, *Rev. Educ. Res.*, **78**(1), pp. 153–189, 2008.
35. K. L. Norman, Spatial visualization – a gateway to computer-based technology, *J. Spec. Educ. Technol.*, **12**(3), pp. 195–206, 1994.
36. M. G. Hewson and M. L. Little, Giving feedback in medical education: verification of recommended techniques, *J. Gen. Intern. Med.*, **13**(2), pp. 111–116, 1998.
37. N. Naikar, R. Hopcroft and A. Moylan, Work domain analysis: Theoretical concepts and methodology (No. DSTO-TR-1665), *Def. Sci. Technol. Organ. Victoria Air Oper. Div*, 2005.
38. K. J. Vicente, *Cognitive Work Analysis – Toward Safe, Productive, and Healthy Computer-Based Work*, CRC Press, 1999.
39. J. Rasmussen, A. M. Pejtersen and K. Schmidt, *Taxonomy for cognitive work analysis*, Roskilde, Denmark: Risø National Laboratory, 1990.
40. F. P. Deek, S. R. Hiltz, H. Kimmel and N. Rotter, Cognitive assessment of students' problem solving and program development skills, *J. Eng. Educ.*, **88**(3), pp. 317–326, 1999.
41. J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, *Eur. J. Sci. Educ.*, **1**(2), pp. 191–203, 1979.
42. J. Larkin, J. McDermott, D. P. Simon and H. A. Simon, Expert and novice performance in solving physics problems, *Science (80-.).*, **208**(4450), pp. 1335–1342, 1980.
43. T. Litzinger, P. N. Van Meter, M. Wright and J. M. Kulikowich, *A cognitive study of modeling during problem-solving*, 2006.
44. K. Wolff, Engineering problem-solving knowledge: the impact of context, *J. Educ. Work*, **30**(8), pp. 840–853, 2017.
45. B. Memarian and S. McCahan, Analysis of Feedback Quality on Engineering Problem-solving Tasks, *Proceedings of the ASEE Annual meeting*, held in Tampa FL, June 16–19, 2019.
46. J. Brooke, SUS-A quick and dirty usability scale, *Usability Eval. Ind.*, **189**(194), pp. 4–7, 1996.
47. S. G. Hart and L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, in *Advances in psychology*, **52**, North-Holland, pp. 139–183, 1988.
48. P. Orsmond, S. Merry and K. Reiling, A study in self-assessment: tutor and students' perceptions of performance criteria, *Assess. Eval. High. Educ.*, **22**(4), pp. 357–368, 1997.
49. M. Yorke, Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice, *High. Educ.*, **45**(4), pp. 477–501, 2003.

Bahar Memarian is a researcher, educator, and analyst with research interests in Human Factors Engineering and Engineering Education. She received her PhD (2021) in Industrial Engineering and the Collaborative Specialization in Engineering Education at the University of Toronto, Canada. Before that, she completed her MASc. (2015) and BASc. (2012) in Electrical Engineering from the University of Toronto.

Susan McCahan is a Professor in the Department of Mechanical and Industrial Engineering at the University of Toronto. She currently holds the positions of Vice-Provost, Innovations in Undergraduate Education and Vice-Provost, Academic Programs. She received her BS (Mechanical Engineering) from Cornell University, and MS and PhD (Mechanical Engineering) from Rensselaer Polytechnic Institute. She is a Fellow of the American Association for the Advancement of Science in recognition of contributions to engineering education has been the recipient of several major teaching and teaching leadership awards including the 3M National Teaching Fellowship and the Medal of Distinction in Engineering Education from Engineers Canada.

Appendix A. Sample problem solution graded using conventional marking and CAIR.

Ideal solution with conventional marking scheme:



Mesh 1 and 2 form a super mesh
Mesh 2 and 3 form a super mesh

Applying KVL to the larger super mesh
$2i_1 + 4i_3 + 8(i_3-i_4) + 4i_2 = 0$ [V]
$i_1+2i_2+6i_3-4i_4 = 0$ [V] [1 mark]

For independent current source, apply KCL to node p
$i_2 = i_1+5$ [A] [1 mark]

For dependent current source, apply KCL to node Q
$i_2 = i_3+3\ i_0$ [A]
$i_0 = - i_4$
$i_2 = i_3 - 3i_4$ [A] [1 mark]

Applying KVL in mesh 4
$2\ i_4 + 8(i_4 – i_3) + 1_0 = 0$ [V] [1 mark]
$5\ i_4 - 4\ i_3 = -5$ [V]
$i_1 = -10$ [A]
$i_2 = -5$ [A]
$i_3 = 40/7 = 5.7$ [A]
$i_4 = 25/7 = 3.5$ [A] $= - i_0$ [1 mark]

Assessment of a sample student solution

- Conventional marking using a numerical grading scheme shown in [*square brackets*]
- CAIR feedback is shown in {*parentheses*} using the tagging codes U, K, T, D, C or A
- An example of the elaboration that could be added is shown in black

$2i_1 + 4i_3 + 8(i_3- i_4) + 6i_2 = 0$
$i_1 + 3i_3 + 6i_3 - 4i_4 = 0$ *[-1/2 mark]* {*A*} Arithmetic error in simplifying first equation
$i_3 = 3i_0$
$i_1 = 5$ *[-1 mark]* {*T*} Missing fundamental understanding of writing constraint equations when there are power sources (e.g., current or voltage) in a branch

$i_0 = - i_4$
$2i_4 + 8(i_4- i_3) + 10 = 0$
$5i_4 – 4i_3 = -5$
$5(-i_0) – 4(3i_0) = -5$
$17\ i_0 = 5$
$i_0 = 0.29$ *[-1/2 mark]* {*D*} Units missing

Total mark: 3/5

Grade on the same sample student solution via CAIR:

Tag incorrect/missing errors on student solution based on alphabetic convention				
	Deep Decomposition		Surface Decomposition	
Goal Abstraction:	**U**nknown variables		**K**nown variables	
Fails○	Below○	Meets○	Exceeds●	
Theory Abstraction:	**T**heoretical model		**D**isciplinary standards within the model	
Fails●	Below○	Meets○	Exceeds○	
Calculation Abstraction:	**C**omputational model		**A**rithmetic work within the model	
Fails○	Below○	Meets●	Exceeds○	