Aligning classroom assessment with engineering practice:
A design-based research study of a two-stage exam with authentic assessment

Milo D. Koretsky¹,² | Campbell J. McColley³ | James L. Gugel³ | Thomas W. Ekstedt³

¹Department of Chemical and Biological Engineering, Tufts University, Medford, Massachusetts, USA
²Department of Education, Tufts University, Medford, Massachusetts, USA
³School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, Oregon, USA

Abstract
Background: Authentic assessment and two-stage exams have recently received attention; however, they are rarely used together. We reimagine assessment by integrating an authentic, computer-based assessment into the structure of a two-stage exam in a large engineering class.

Purpose: We seek to identify ways that such assessment extends classroom testing to better align with engineering practice by examining the ways teams negotiate uncertainty to make engineering decisions. We also identify differing students’ reactions to increased uncertainty during tests.

Design/Method: Using the methodical framework of design-based research, we analyze performance and reflection data for 117 student teams through two design iterations to explore four design and theoretical conjectures.

Results: Teams chose multiple solution paths to this authentic task, an aspect that aligns with the characteristics of engineering practice that we seek to assess. In addition, the technology tool allows the evaluation of procedural accuracy for many of the teams’ chosen paths. The teams’ decision-making performances correlate; however, decision-making and traditional assessments do not correlate, suggesting they measure different competencies. The computer-based second stage provides a holistic assessment that shifts the messages that students implicitly receive about valued practices in the classroom. However, not all students took up the authentic group assessment in desired ways.

Conclusions: Technology-based two-stage exams with authentic assessment show promise to shift testing practices in large engineering classes to include decision-making. Such assessments better align with engineering practices that are valued in the profession, but more work is needed to develop systems for widespread implementation.

KEYWORDS
chemical engineering, design-based research, higher education, learning technology, problem-solving, student assessment

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. Journal of Engineering Education published by Wiley Periodicals LLC on behalf of American Society for Engineering Education.

J Eng Educ. 2022;111:185–213.
1 | INTRODUCTION

Practicing engineers are assessed based on their ability to get the job done (Vincenti, 1990)—that is, develop and deliver a solution that satisfies a set of objectives and competing constraints. Engineers’ work involves ambiguity where they often make decisions based on incomplete and uncertain knowledge, and this aspect leads to many possible solution paths (Murray et al., 2019; Trevelyan, 2014; Vincenti, 1990). In contrast, assessment practices in schools commonly place students in unrealistically limited contexts (Biggs, 2014; Boud, 1990; Jordan & Babrow, 2013). For example, they are asked to solve problems alone, allowed limited resources, and problems are designed to be deterministic and have one clearly appropriate solution path. In part, instructors design school-based assessments this way as a matter of practicality, especially when teaching large enrollment classes. They want the assessments to be fair and without a large time burden to grade.

In the design-based research (DBR) study reported here, we investigate how the use of a computer simulation in a two-stage exam (Efü, 2019) was able to align student assessment with engineering practice, an approach termed authentic assessment (Villarroel et al., 2020; Wiggins, 1990). Our approach builds upon identified differences between the nature of professional engineering practice and the work that students complete in many engineering classrooms. As Jonassen et al. (2006) state, “Workplace engineering problems are substantively different from the kinds of problems that engineering students most often solve in the classroom; therefore, learning to solve classroom problems does not necessarily prepare engineering students to solve workplace problems” (p. 139). In part, this difference arises as students are not asked to frame problems and make decisions (Douglas et al., 2012). In the assessment design reported here, the first exam stage contains a traditional individual assessment. In the second stage, student teams use a technology tool to complete an authentic, industrially situated engineering task. The second stage, therefore, elicits collaborative decision-making. At the same time, this two-stage exam fits within the instructional practices of many large undergraduate engineering classes.

The DBR study sits within a broader organizational change initiative in a 4-year engineering department. The initiative aims to shift learning activity to more closely resemble professional practice by using problem-based learning activities within a studio structure (Koretsky, Keeler, et al., 2018). Revised studio problems position students in the role of engineers on teams where they need to identify core foundational principles as conceptual tools to progress on tasks that resemble realistic engineering work (Engle & Conant, 2002; Johri & Olds, 2011). In a previous study during this initiative, we used an activity systems framework (Boaler & Greeno, 2000; Engeström, 2001) to reveal elements that influence students’ adoption of more or less productive approaches to learning. By far, exam expectations were the most common element cited by students to prompt undesired rote learning approaches (Michor & Koretsky, 2020). The two-stage exam with authentic assessment reported here was developed to change students’ exam experiences and thereby influence the ways students approach their learning.

This study reports two design iterations where a two-stage exam was used in a large (~200 students) undergraduate engineering science course in an ABET-accredited program in the United States. Using the technique of conjecture mapping (Sandoval, 2014), we explore the following design conjectures (DC) and theoretical conjectures (TC):

DC1. Identifying teams’ engineering decisions provides information about ways students are able to use core knowledge in practice.

DC2. Technology-based assessments can automate the evaluation of procedural accuracy across different data sets and solution paths.

TC1. Authentic assessments measure teams’ problem formulation and decision-making competencies in ways not available in typical classroom assessments.

TC2. Students react to shifts in assessment practices based on their encultured ideas about schooling and testing.

In the following sections, we provide the background and theoretical framework for the study of the second stage’s computer-based authentic assessment and describe features of the technology tool. We then present findings and discuss each of the research conjectures.

2 | BACKGROUND AND THEORETICAL FRAMEWORK

Assessment plays a key role in the engineering classroom. Assessments are used to certify mastery and determine grades and also to provide educators information about the effectiveness of instruction (Johnson & Johnson, 1999;
Pellegrino et al., 2001). Importantly, they also indicate to students the valued knowledge, skills, and practices in the discipline (Boud, 1990; Hargreaves, 1997). In this section, we argue that computer-based assessment tasks can place students in the role of engineers doing realistic work. This authentic assessment aligns with a practice-centered perspective of learning (Manz et al., 2020) where managing uncertainty to make decisions is an important skill for engineers. Finally, we discuss how we adapted the two-stage exam structure to accommodate such an authentic assessment within the constraints of a large enrollment class in engineering.

2.1 Technology-enhanced assessment

Technology-based assessments are increasingly being integrated into educational contexts (Bearman et al., 2020; Koretsky & Magana, 2019; Thomas, 2016). These assessments can take many forms including the use of data banks of multiple-choice questions with automated marking and feedback (B. Chen, West, & Zilles, 2019; Ćukušić et al., 2014; Mitkov et al., 2006), the use of learning analytics to increase student success (Larrabee Sønderlund et al., 2019), and embedded assessments immersed in serious games (Caballero-Hernández et al., 2017; Kim & Shute, 2015). Supporters see the potential of technology to enable learning environments that provide real-time feedback and scalable and personalized support. For example, randomized computer-generated “rolling problems” have been used to allow students to take exams asynchronously in large enrollment courses (Sud et al., 2019).

However, others have critiqued common uses of these technologies stating they overemphasize practical instructional objectives such as efficiency, ease of grading, and record keeping (Wiggins, 1990). Such approaches often use highly specific and structured questions that form piecemeal assessments where instructors reduce a topic into several separate and discrete parts, assess their students’ performance on each, and determine a score by adding those together (Haladyna et al., 2002). This type of assessment can reward rote learning: the memorization of facts and solution algorithms (Pellegrino & Quellmalz, 2010). Biggs (2014) advocates for instructors to replace piecemeal summative assessments with assessments that holistically address the “whole performance” (p. 8). Boud (1990) elaborates that those piecemeal assessment practices are antithetical to both the disciplinary practices instructors utilize in their own scholarly work and to the broad goals of university education.

In this spirit, educators argue that technology-based systems present an opportunity to reimagine assessment (Behrens et al., 2019; Bucciarelli, 2003; Timmis et al., 2016). They advocate for a paradigmatic shift from viewing the affordances of technology as primarily toward the efficient delivery at scale to rethinking the fundamental relationship between learning and assessment. Toward this end, the use of immersive simulations to create scenarios where students are placed in the role of practicing professionals doing complex and meaningful work becomes compelling (Barab et al., 2007; Clark et al., 2009; Quellmalz & Pellegrino, 2009; Siyahhan et al., 2017). Such technology-based simulations have been used to assess important engineering competencies like open-ended problem-solving, critical thinking, and modeling and design skills (Koretsky et al., 2008; Rupp et al., 2010; Vieira et al., 2016; Xie et al., 2014; Xing et al., 2020). Thus, the core skills that are assessed shift from answering questions that only have value in a school context to doing meaningful science or engineering work. However, such an approach also has risks. In game-based simulations, the assessment component may be tacit, leading some learners to be less careful about details as they complete their work (Winkley, 2010).

The second stage of the two-stage exam studied here builds on several areas that Timmis et al. (2016) identify to reimagine assessment, including using technology-enhanced assessment to provide learners complex decision-making opportunities where learners can work collaboratively and exercise more agency than with traditional assessment practices. This approach shifts the assessment activity from piecemeal practices to completing an engineering task with realistic goals. At the same time, the second stage is identified clearly as being part of an exam that students should seriously engage in.

2.2 Situated learning

Assessment practices are a powerful indicator to students of what is important to learn and, therefore, should be fundamentally grounded in principles and theories of learning (Hattie & Brown, 2007; Timmis et al., 2016). The computer-based assessment design studied here draws on the work of learning scientists over the past 30 years, which shifts the view of learning from a cognitive approach that aligns with an acquisition metaphor to a practice-centered
approach that is more appropriately described through legitimate participation (Barab & Duffy, 2000; J. Brown et al., 1989; Engle & Conant, 2002; Lave & Wenger, 1991; Manz et al., 2020). According to the practice-centered perspective, knowing and doing are intertwined, that is, what is learned is not separate from how it is learned. Learning depends on the content, context, and activity, and knowledge is situated in the experience (Barab & Duffy, 2000; Dewey, 1938; Turner & Nolen, 2015). This view fundamentally challenges the notion that concepts are self-contained entities but rather positions concepts as tools, which can only be fully understood through use. Thus, learning involves more than “acquiring” conceptual understanding, but rather involves having students build an “increasingly rich implicit understanding of the world in which they use the (conceptual) tools and of the tools themselves” (J. Brown et al., 1989, p. 33). This understanding is framed by those situations in which the conceptual tools are learned and used. In the context of the computer-based engineering assessment developed here, this perspective implies that an assessment tool should not only provide a set of tasks on which to evaluate a student but should also embed that assessment within a context where those tasks have a meaning that is connected to the engineering profession and its practices.

2.3 | Uncertainty and decision-making in engineering practice

Making progress in the face of uncertainty is fundamental to engineering and science practice (Pickering, 1995; Trevelyan, 2014; Vincenti, 1990). We follow Manz and Suárez's (2018) definition of uncertainty as to the aspects of an engineer’s work that is “non-obvious and contingent, which must be figured out by the scientist [or engineer] and negotiated in response to feedback from peers and the material world” (p. 288). In both engineering contexts (Douglas et al., 2012; Jonassen et al., 2006) and science contexts (Y. C. Chen, Benus, & Hernandez, 2019; Manz, 2015), researchers have identified fundamental differences between how uncertainty manifests in school and practice. In school, students are encultured to knowledge that is certain as measured by assessments with unambiguous solution paths. From this perspective, uncertainty is viewed as undesirable and effort should be taken to avoid it (Jordan & Babrow, 2013). In contrast, real engineering work contains knowledge that is initially unknown, is legitimately problematic, and only temporally emerges with experience and understanding (Manz et al., 2020; Pickering, 1995). By struggling to negotiate uncertainty and building on one another’s perspectives to resolve it collaboratively, learners develop a deeper understanding of core scientific principles, even if their immediate resolution may not be normatively correct (Jordan & Babrow, 2013; Lotan, 2003; Manz, 2015).

In the context of assessment, we argue that uncertainty is an important element of an authentic task, especially for group work. From this perspective, assessment in engineering should address the ability to work with others to navigate uncertainty. One way such uncertainty is manifest is by problems with many distinct potential solution paths. Borrowing from Jonassen’s (2000) typology, the task studied here is constructed as a decision-making problem where different engineering decisions can lead to different solution paths. Such problems involve identifying benefits and limitations, weighing options, selecting an alternative, and justifying those choices (Jonassen, 2000). While decision-making problems do not embody the complete extent of uncertainty that professional engineers face, they do extend students further into authentic practice than they commonly experience in their engineering exams.

2.4 | Authentic assessment

Equipping students with the skills to use their knowledge in professional practice is an important goal of many university programs. Correspondingly, the role of assessment is appropriate to measure how much learning has occurred and to provide feedback on the effectiveness of instruction (Johnson & Johnson, 1999; Pellegrino et al., 2001). Equally important but often less recognized, assessments also orient students to the nature of knowledge and the legitimate ways that knowledge is used in the profession (Biggs, 1996; Boud, 1990; Hargreaves, 1997; Wormald et al., 2009). The types of knowledge that students develop and the ways they learn to use it are mediated by the assessments by which they are evaluated. Therefore, assessments become critical in orienting students to take up desired engineering practices. As stated earlier, a key motivation for the shift in assessment studied here was based on student reports that traditional course assessments prompted undesired rote learning practices.

Authentic assessment uses real-world tasks that mirror the work required of professionals, in our case engineers (Villarroel et al., 2020; Wiggins, 1990). This type of assessment requires students to engage in the type of reasoning and problem-solving needed in their field. Thus, students must demonstrate deep understanding, higher-order thinking, and
complex problem-solving. Importantly, authentic interactions should also mirror the social interactions of engineering practice. Characteristics of authentic assessments include problem context, characteristics, and student–student interactions (Jonassen, 1997). The assessment should be set in a context that reflects how students will use their knowledge in the future and include self-reflection (Villarroel et al., 2020). Problems should be realistic, partly ill-defined (requiring the students to frame the problem), and complex with multiple solution paths possible. The work should be collaborative with students working in groups and other students’ perspectives proving valuable resources (Bucciarelli, 2002; Lotan, 2003).

2.5 Two-stage exams

In this study, we integrate the concept of authentic assessment into the structure of a two-stage exam. Recently, two-stage exams have gained attention as an assessment practice that can simultaneously improve learning (Brame & Biel, 2015; Efu, 2019; Rieger & Heiner, 2014; Stearns, 1996; Zipp, 2007). In the first stage, students complete a traditional exam individually; in the second stage, students work collaboratively in groups. Student scores are determined by a weighted average on both parts. Similar to the way students’ collaboration in peer instruction builds understanding when concept questions are used in formative assessment (Mazur, 1997), during the second stage, collaborative sense-making can lead to improved understanding during summative exams. Importantly, group exams have the potential to assess social competencies needed in the professional workplace (Jang, 2016) and send students the message that collaboration is a valued central practice of doing science and engineering (Boud, 1990; Gilley & Clarkston, 2014; James, 2014; Rieger & Heiner, 2014). However, while two-stage exams appear to improve motivation and lower stress, there are mixed reports about whether they actually improve learning (Gilley & Clarkston, 2014; Kinnear, 2020; Leight et al., 2012).

Most recent reports use two-stage exam designs that ask student groups to solve the same constrained problems in the second stage as they did as individuals in the first stage. However, such an approach may not provide the fullest opportunity to leverage their team members’ different ideas, perspectives, and competencies. Indeed, educators have argued that problems that are “context rich” (Heller & Hollabaugh, 1992) or “group-worthy” (Lotan, 2003) provide tools that better align with the ways knowledge and skills are used by practicing professionals. Similarly, these problems can provide the basis for authentic assessment. In the assessment design reported in this study, we build on the idea of two-stage exams by basing the group portion of the second stage on a more authentic group-worthy task: a task enabled by a computer simulation as described in the next section.

3 Industrially situated technology design

3.1 Task design

The situated task for the second stage places students in teams as engineers where they must determine the rate constant for a sugar reaction and use the rate constant results to recommend a reactor run time for an industrial candy manufacturing process. The industrial context is provided by a 3D computer simulation (see Figure 1). Students navigate this environment in “first-person” mode, engaging with equipment and instruments to run the reactors (left) and take measurements (right). After analyzing the results in collaboration with teammates, each individual student submits responses (see Figure 2). To support the context, the assignment (see Appendix A) is in the form of a memorandum from the Vice President of Engineering, and teams are provided an “equipment manual” that describes procedures to run the reactors and take measurements. They are also provided the rubric (see Appendix B) with guidelines for collaborative group work by which they will be assessed.

The simulation is built on a WebGL software platform and runs in standard web browsers. Reactor data are produced by a kinetic model with added random noise. The output provided by the kinetic model affords teams the opportunity for scientific reasoning, while added variation (noise) provides opportunity for statistical reasoning. Each team collects data from two reactors, and the simulation generates unique results for each team member. In effect, this feature provides the same function as randomized computer-generated “rolling problems” as one team’s reactors performance is fundamentally different from their classmates. In some cases, the two reactors have the same kinetic parameters, meaning that outside the process and measurement noise they will perform the same, while in other cases the two reactors are inherently different. The simulated reactor time–concentration data and the students’ submitted responses are stored in a central database.
3.2 | **Iteration of technology design**

We implemented two changes in the simulation for the second iteration of data collection based on analysis of the previous cohort’s data. First, we enhanced the technology design to support team collaboration by enabling each student to view and access the results of their teammates along with their own results, as shown in Figure 3. Each student...
independently engages in the situated industrial environment; the intent is for them to each be able to collect data and then combine data with their teammates into a superset that they analyze collaboratively. Analysis of the previous cohort revealed that 59 students (27%) worked individually or reported analysis based only on their individual data set. After the redesign shown in Figure 3, only five students (3%) worked as individuals. To compare across cohorts, we use only data from the students whose submissions are consistent with collaborative work.

Second, we altered the first question of the submission form to the one shown in Figure 2. For the previous cohort, question 1 had a response box the same size as question 2. This change was made as a cue to encourage teams to report interval estimates rather than point estimates, while not explicitly directing them to do so.

4 | METHODS

4.1 | Methodological framework

The study uses the methodological framework of DBR in education. In DBR, innovative educational systems are deployed in naturalistic settings while, simultaneously, experiments studying the innovative systems are systematically conducted (A. L. Brown, 1992; Collins, 1992; Sandoval, 2014). Through iteration, instructional designs are improved and theoretic conjectures are refined, that is, design of instruction and research on learning intermingle (Bakker, 2018; Minichiello & Caldwell, 2021). Like engineering design, itself, DBR uses a systems approach to create useful innovations within real constraints (Hjalmanson & Lesh, 2008; for examples of DBR studies in engineering, see Dasgupta, 2019; Newstetter, 2005). Importantly, just as Vincenti (1990) argues that the engineering design process produces new theoretical knowledge unique to engineering, DBR can provide new theoretical knowledge unique to learning and instruction. Indeed, Hjalmanson and Parsons (2021) argue that the epistemological differences between DBR and experimental or quasi-experimental education studies are akin to the fundamental differences between engineering and science.

Kelly (2004) describes a method’s argumentative grammar as “the logic that guides the use of a method and that supports reasoning about its data” (p. 118). DBR operates within specific classroom learning ecologies that mandate a
different grammar than experimental or quasi-experimental designs (Cobb & Gravemeijer, 2008; Reimann, 2011). Sandoval (2014) developed the technique of conjecture mapping as a tool to articulate the argumentative grammar of a DBR study, that is, to identify the salient theoretical features of a learning environment and predict how they interact to produce desired outcomes.

Figure 4 illustrates our adaptation of Sandoval’s (2014) conjecture mapping to relate two high-level conjectures— theoretically grounded ideas of how a technology tool can support desired assessment of learning—to more specific design and theoretical explorations pursued in this study. These high-level conjectures are embodied within the tools and materials, task structures, and participant structures of the assessment process. This embodiment leads to system-specific design and theoretical conjectures that connect data sources and data analyses to support empirically grounded claims.
4.2 | Positionality

We (the four authors) are US born, able-bodied, cisgender White males, three straight and one queer, who all have engineering degrees. We bring differing perspectives to the technology development, instruction, and research described here. Two of us have extensive industrial experience. The first author directs a research program that seeks to develop learning systems to allow students to be able to integrate and extend the knowledge developed in specific courses in the core curriculum to the more complex, authentic problems and projects they face in professional practice. He has previously conducted DBR studies of technology systems targeting professional practice (Koretsky et al., 2011) and connected conceptual understanding (Friedrichsen et al., 2017) and of organizational change initiatives focusing on higher-level cognitive and social skills in engineering problem-solving (Koretsky, Keeler, et al., 2018; Koretsky, Montfort, et al., 2018). The second author is a PhD student in environmental engineering. He brings interests in both technical and educational research to the project. His technical interests center on microplastic contaminant fate and transport in surface waters, while his educational interests target difference, power, and discrimination in environmental engineering. The third author participated in this project as an undergraduate student pursuing a BS in bioengineering. He completed the course studied here the year before the two-phase exam was implemented and brought perspectives of an enrolled student. He has graduated and currently works as a brewer for a local microbrewery. The fourth author develops software solutions that support STEM education research in the areas of conceptual understanding, integrated learning, and active learning systems. Previously, he has worked in a variety of industrial settings developing IT systems that support manufacturing processes.

As privileged White male engineers, we acknowledge the limitations of our collective experience. This work should be considered with that limitation in mind.

4.3 | Participants and setting

This study was conducted at a public, research-intensive university in the Pacific Northwest region of the United States. The two-stage midterm exam was delivered in a large course required for biological, chemical, and environmental engineers. Data are reported from two cohorts during the second and third DBR iterations in consecutive spring terms. All enrolled students were invited to participate. In Cohort 1, 213 students consented to participate in the study, and in Cohort 2, 168 students consented. Students were mostly sophomores and juniors. Table 1 shows self-reported demographic data. The research was approved by the institutional review board.

Figure 5 shows a flowchart of class activity through the first 6 weeks, highlighting the assessment in Week 6. In Weeks 1 through 5, students participated in regular weekly class activity, which included two 2-h lectures on Tuesday

---

**Table 1** Self-reported demographic data by cohort based on a 69% response rate from study participants

| Classification | Category          | Cohort 1 (%) | Cohort 2 (%) |
|----------------|------------------|--------------|--------------|
| Major          | Biological       | 22           | 26           |
|                | Chemical         | 70           | 64           |
|                | Environmental    | 8            | 10           |
| Gender         | Female           | 25           | 31           |
|                | Male             | 73           | 66           |
|                | Non-binary/another gender | 2 | 3 |
| Race           | Asian            | 14           | 18           |
|                | Black            | 3            | 4            |
|                | Hispanic/Latinx  | 6            | 9            |
|                | Pacific Islander | <1           | <1           |
|                | White—non-Hispanic | 61     | 56           |
|                | International    | 11           | 7            |
|                | Two or more races | 7        | 6            |
and Thursday, which the entire class attended. Lectures were interactive and utilized an audience response system called the Concept Warehouse (Koretsky et al., 2014). To prepare for lecture, students were assigned pre-reading and asked to answer questions in the Concept Warehouse related to the reading assignment. Each student also attended a 2-h studio section with a maximum enrollment of 24 where they completed activity-based group work facilitated by a graduate student and an undergraduate student. The majority of the teams had three members (87%), with the remaining teams consisting of two or four members. The team composition for the second stage of the midterm exam was the same as that for the first 5 weeks in studio. Weekly homework assignments were submitted in Weeks 2 through 5 before the midterm exam.

The two-stage midterm exam was delivered during the sixth week of the 10-week term. Students completed the first stage, the individual portion, during lecture on Tuesday. They were allocated the entire class period and allowed to bring only a single $8 \times 11$" note sheet and a non-graphing calculator as resources. Five members of the instructional team circulated around the room to proctor and answer questions of clarification. Approximately half of the students finished within 90 min. Collected exams were scanned into Gradescope (a computer-based grading tool) and then graded by the instructional team. Appendix C provides a sample item that assesses conceptual understanding and one that assesses procedural accuracy. The other items are similar and not presented for brevity.

Student teams completed the second stage, the group portion, the following day during studio. A graduate and an undergraduate student instructor observed the teams and evaluated their engagement and collaboration using a rubric provided by the class instructor with explicit guidelines for collaboration (Appendix B). Students were also assessed on their submitted solutions (Figure 2). The second stage was “open everything but other humans.” Team members used laptops to access the simulation, perform analysis in MATLAB or Excel, and could access course content and conduct more general internet searches. Most teams used the entire 2-h allocation to complete this portion.

Instead of the usual reading pre-quiz for Thursday’s lecture, students answered a brief reflection about the midterm. The class period was used to interactively explore the different possible solution paths and to reflect on group teaming strategies. At the end of Thursday’s class, students were asked another reflection question. The question for Cohort 2 was different than for Cohort 1.

### 4.4 Data sources

Student performance and reflections formed the basis of the data analyzed in this study. These data sources relate to the explored design and research conjectures, as shown in Figure 4. The analysis of performance data focused on the second stage. The technology output from the sugar reactor task consisted of the following information for each
student: their team numbers; the unique reactor parameter values provided to each team; the raw simulated data pro-
vided to each team member; and individuals’ responses to the four questions in Figure 2. The responses allowed us to
infer the decisions leading to their solution path and to assess their procedural accuracy by determining if the calcula-
tions were correct. Students also submitted supporting computational artifacts such as Excel and MATLAB files. Grades
from the first stage of the midterm and the final exam were available from a technology-based grading tool called Grad-
escape. Student scores for the Concept Warehouse audience response system also provided a measure of student
performance.

Students responded to a set of reflection prompts after the two-stage midterm exam, including free-response items
and Likert scale items. After reading the three sets of written responses, the research team selected the following
prompt for detailed coding of student perceptions:

Exam Vent: This week was busy, with an in-class individual Midterm and a group problem in studio. Some-
times it is useful to vent. In the space below, please provide any comments you feel like on these
assessments.

We selected this prompt as it addressed both stages and was framed in terms of “venting,” which could probe ways
that the exam disrupted expectations (theoretical conjecture 2). The other two prompts (write down one thing you
learned from the Group Midterm; was there anything you could have done differently to help your group collaborate
towards the Group Midterm solution?) were used for triangulation of the detailed coding results.

4.5 | Data analysis

4.5.1 | Student performance

The sugar reactor task is designed for teams to toggle between engineering and statistics principles as they complete
their work. Teams need to work collaboratively to organize their approach and make engineering decisions. While each
team made many choices during the task, we analyzed the following major decisions that were central to most teams’
approaches:

1. Mathematical form: How did they mathematize the kinetic rate expression?
2. Solution strategy: How did they organize the data collected from two reactors and several team members to perform
   their analysis?
3. Determination of rate constant: Did they provide a point estimate (single value) or an interval estimate (confidence
   interval) for the first-order rate constant?
4. Time calculation: Did they account for process variation in selecting a process time for manufacturing? How?
5. Statistical analysis: Did they determine if the two reactors’ kinetic response was statistically different?

The unit of analysis for student performance is the team. As described above, we only analyzed data from teams
who submitted consistent values across at least two team members, indicating they worked collaboratively. Data from
53 valid teams from Cohort 1 and 64 teams from Cohort 2 were coded. Averages of team members’ scores to individual
assessments (Final exam, Stage 1, audience response system) were used for correlational analyses.

The technology tool includes a set of prompts (Figure 2) in which students report their findings and engineering
recommendations and justify them. We first coded the submissions to infer choices for the decisions identified above.
The decision codes, mathematical form, solution strategy, determination of rate constant, time calculation, and reactor
difference, are described in Table 2. These codes were developed during the first DBR iteration by examining the initial
results and feedback from students during the interactive, in-class reflection activities the following day. These codes
remained fixed, and data from the second and third DBR iteration were reported. Using a set of key words for each
code together with student values and explanations of the rate constant and production times, researchers identified
each team’s set of decisions. We counted a decision only if it was incorporated into their submitted recommendation;
for example, if a team identified they would calculate a confidence interval but ran out of time, it was not coded as
“confidence interval.” Two raters independently analyzed 40 responses, selected at random. Fleiss's kappa was used to
evaluate interrater reliability. Kappa values for mathematical form (0.906), solution strategy (1.000), confidence interval
(0.950), and time calculation (0.922) indicate “almost perfect agreement” according to Landis and Koch (1977). All discrepancies were resolved by consensus. As a check, code results for 10 teams with different solution paths were compared with their submitted Excel or MATLAB supporting files. In all cases, the two matched. A single rater coded the remaining responses.

We then assessed a team’s procedural accuracy as follows. The database from the technology tool stored each team member’s raw data. Based on a team’s selected identified solution path, we calculated the values for rate constant and manufacturing time and compared those calculations with the team’s submitted values. We used the team’s calculated values for the rate constant determination to determine the correctness of their time calculation; thus, if they did not get the former correct, points did not get deducted twice. In this procedural assessment, there were several additional team decisions that we needed to identify, including if a team chose to analyze the two reactors separately and what significance level they chose to use (most chose \( \alpha = 0.05 \), but some chose .01). If the submitted values agreed to our calculations within round off, it was labeled as correct. We were not able to determine correctness for the decisions coded “Other.”

### 4.5.2 Student reflections

The unit of analysis for student reflections is the individual. Individual student-written responses to the reflection question asked at the end of Thursday’s class for Cohort 1 (Exam Vent) were analyzed using emergent coding and thematic analysis (Braun & Clarke, 2006; Riessman, 2008). After a stable set of code themes was established, axial coding was conducted by the second author (Strauss & Corbin, 1990).

To develop the initial code sets, approximately 20 responses were coded collaboratively by the first, second, and fourth authors. The second author then coded the remaining responses, identifying issues with the codes and missing
elements. The team then met to reconcile the code set. This process was repeated until the code set no longer changed. After a stable set of code themes was established, the second author coded the entire set.

This coding process led to two sets of code categories. The first set (Table 3) focused on the content of the response and included the nature of the assessment problem, expectations of assessment, social context/group work, and value of experience. The second set (Table 4) was based on how the responses aligned with the intention of the assessment. Code categories included encultured in schooling, situated in engineering, and hybrid. When the initial set of coding was complete, the first and second authors read through the other reflective questions with an eye for confirming or disconfirming evidence and modified the summary results accordingly.

5 | RESULTS

We present results to address the four research conjectures. We first characterize the solution paths of teams across the two cohorts showing teams make different decisions as they grapple with the uncertainty of the task. We then present the procedural accuracy of the teams’ submitted results and show correlations between these performances and traditional assessments. Finally, we characterize student attitudes through their reflections, showing that some students did not engage in authentic group assessment in ways that we intended.

5.1 | Team performance

5.1.1 | Design conjecture 1: Engineering decisions

Figure 6 shows decision maps for the solution paths for Cohort 1 (Figure 6a) and Cohort 2 (Figure 6b) by characterizing four decisions: mathematical form, solution strategy, determination of rate constant, and time calculation. The width of the lines connecting decisions is proportional to the number of teams making that choice. For example, in Cohort 1, the highest proportion of teams that used a linear representation of the kinetic model used a “k's per Reactor” solution strategy; however, some teams chose each of the other three solution strategies.

In general, the teams followed many different solution paths in the sugar reactor task based on their collaborative decision-making. A Kruskal–Wallis test shows the decision proportions for mathematical form ($H = 0.098; p = .75$) and time calculation ($H = 0.024; p = .88$) generally align among cohorts and do not show statistical difference.
Cohort 2 selected less $k$'s per time (solution strategy; $H = 17.8; p < .001$) and more often decided to report an interval estimate than a point estimate ($H = 13.0; p < .001$). These choices align with desired engineering decisions. The latter change is also consistent with the second technology improvement described above (changing the size of the response box). Many teams labeled as “Other” for the time calculation recognized they needed to account in the observed variation in the experimental runs, but rather than using a statistical model to quantify that variation and make a resulting recommendation, they simply looked directly at the raw data from each run and used the slowest of that limited set to roughly estimate a recommended time.

**FIGURE 6** Sankey diagrams showing the solution paths mapped for (a) Cohort 1 ($n = 53$ teams) and (b) Cohort 2 ($n = 64$ teams)
We present data on teams’ decision to use a statistical test to compare reactors in Table 5. While some teams had reactors that were inherently different, others did not. This decision was not included in Figure 6 as the two cases are not symmetric. It is possible a team whose reactors were not different, performed the analysis and simply did not report it. On the other hand, teams whose reactors were different would more likely report it if that was investigated properly. Few teams whose reactors were different reported such an analysis (5 of 55 teams), even though they completed homework problems, a studio activity, and a Stage 1 exam question on statistical comparison for two samples. One team whose reactors had different rate kinetics explained they used Reactor 1 as it was slower but did not report doing statistical test of reactor difference. Thus, while many students demonstrated proficiency with the procedural calculation in traditional assessments, it did not occur to them to use this approach strategically when they needed to make an engineering decision in Stage 2. In addition, 10 of 62 teams whose reactors were not different chose to report rate constants for each reactor separately. Several of these teams reported confidence intervals that clearly overlapped.

We can use these data to see the extent to which teams made decisions that align with appropriate engineering practice. Exactly 25% of teams in Cohort 1 and 59% in Cohort 2 chose to use an interval estimate rather than a point estimate to report the rate constant. Only 37% (Cohort 1) and 30% (Cohort 2) of teams used statistical modeling to account for variability in their time recommendation to manufacturing (CI_Low). In practice, not accounting for variability would lead to many batches in production not meeting specification. Finally, for the cases where the two reactors were different, less than 10% of teams reported performing a statistical analysis of difference.

### 5.1.2 Design conjecture 2: Procedural accuracy

Table 6 shows the calculation accuracy for each cohort for each of the two main deliverables, according to code category, with the other decision choices aggregated. The teams that chose an interval estimate for the rate constant determination or time calculation had a lower percentage correct than the more straightforward point estimate calculations. In several cases, we could identify a specific mistake a team made. For example, within the “$k$’s per Reactor” decision for solution strategy, seven teams used a $z$-value rather than the appropriate $t$-value to calculate the confidence interval. A few groups using this strategy included five runs in their average instead of all six and consequently did not report $k$ correctly. Conversely, there were a few teams with legitimate creative approaches that were “outside the box” and needed to be evaluated accordingly. It would be challenging to automate the process for computer grading in such cases.

The percentage correct reported in Table 6 are comparable to the scores on the traditional Stage 1, which assessed procedural accuracy and conceptual understanding, but not engineering decision-making (75.8 ± 11.0% [Cohort 1]; 68.4 ± 13.6% [Cohort 2]).

### 5.1.3 Theoretical conjecture 1: Student competencies

Table 7 shows the results of a Pearson’s correlation analysis among components of Stage 2 assessment including decision-making and procedural accuracy (calculation) for rate constant and time with traditional summative (Final exam, Stage 1) and formative (audience response system) assessment measures. In this analysis, the scores for the assessment measures of individual team members (Final exam, Stage 1, and audience response system) were averaged.
across the team to be consistent with the team data in Stage 2. The decision-making components positively correlate to each other and the traditional assessments positively correlate to one another. However, the components of Stage 2 do not correlate to any of the traditional assessments. This result indicates the individual summative (Final exam, Stage 1) and formative (audience response system) assessments measure different competencies than the group decision-making assessment (Stage 2). We infer from these correlations and the results from Figure 6 that the authentic assessment reported here measures teams’ problem formulation and decision-making competencies in ways not available in typical classroom assessments.

The correlation of these assessment measures between cohorts is also shown. Only decision-making measures correlate with cohort. This finding indicates the two technology changes described above may have supported the teams’ decision-making, providing support for the iterative DBR approach reported here as a means for improvement of educational technology.

### TABLE 6
Student procedural performance for different solution paths

| Deliverable       | Rate constant determination | Time calculation |
|-------------------|-----------------------------|------------------|
|                   | Confidence interval         |                  |
|                   | Code                        |                  |
|                   | Cohort 1        | Cohort 2        | Cohort 1        | Cohort 2        | Cohort 1        | Cohort 2        |
| Correct           | 6              | 19              | 28              | 18              | 11              | 15              |
| Incorrect         | 7              | 18              | 12              | 9               | 8               | 4               |
| Total             | 13             | 37              | 40              | 27              | 19             | 19              |
| Percent correct (%) | 46             | 51              | 70              | 67              | 58             | 79              |

### TABLE 7
Pearson’s correlation analysis of group assessments, individual assessments, and cohort

| Decision rate_k | Calculation rate_k | Calculation time | Final exam | Midterm Stage 1 | Audience response system | Cohort |
|-----------------|-------------------|------------------|------------|-----------------|--------------------------|--------|
| Pearson         | .4693*            | −.1570           | .1448      | .0989           | .1262                    | .1498  | .3309* |
| correlation     |                   |                  |            |                 |                          |        |        |
| p value         | .0000             | .1030            | .1330      | .3060           | .1908                    | .1199  | .0004  |

| Decision time   | Calculation rate_k | Calculation time | Final exam | Midterm Stage 1 | Audience response system | Cohort |
|-----------------|-------------------|------------------|------------|-----------------|--------------------------|--------|
| Pearson         | −.0568            | .1332            | −.0861     | .0042           | .0786                    | .3853* |
| correlation     |                   |                  |            |                 |                          |        |        |
| p value         | .5576             | .1675            | .3732      | .9651           | .4163                    | .0000  |

| Calculation rate_k | Calculation time | Final exam | Midterm Stage 1 | Audience response system | Cohort |
|-------------------|------------------|------------|-----------------|--------------------------|--------|
| Pearson           | .1686            | −.0006     | .1309           | .0430                    | −.0976 |
| correlation       |                   |            |                 |                          |        |        |
| p value           | .0797            | .9951      | .1748           | .6572                    | .3128  |

| Calculation time | Final exam | Midterm Stage 1 | Audience response system | Cohort |
|------------------|------------|-----------------|--------------------------|--------|
| Pearson          | −.1094     | .0216           | −.0730                    | .0904  |
| correlation      |            |                 |                          |        |        |
| p value          | .2575      | .8239           | .4506                    | .3499  |

| Final exam | Midterm Stage 1 | Audience response system | Cohort |
|------------|-----------------|--------------------------|--------|
| Pearson    | .6453*          | .4127*                   | −.0819 |
| correlation|                 |                          |        |        |
| p value    | .0000           | .0000                    | .3970  |

| Midterm Stage 1 | Audience response system | Cohort |
|-----------------|--------------------------|--------|
| Pearson         | .4621*                   | −.0679 |
| correlation     |                          |        |        |
| p value         | .0000                    | .4832  |

| Audience response system | Midterm Stage 1 | Cohort |
|--------------------------|-----------------|--------|
| Pearson                  | −.0430          |        |
| correlation              |                 |        |        |
| p value                  | .6573           |        |

*Denotes significant correlation at p value < .01.
Theoretical conjecture 2: Student assessment experience

We next summarize the main findings from the analysis of student reflection responses to the “Exam Vent” prompt. We observe two points of view in students’ responses, encultured in schooling and situated in engineering. Table 8 presents sample quotations for each category with one column representative of encultured in schooling and the next situated in engineering. Responses coded as encultured focused on school work and the importance of grades, whereas those coded as situated were reflective of doing real engineering work and often connected to engineering identity or understanding. Not all students engaged in the group stage in the spirit of an authentic engineering task.

Table 8 also shows the number of responses (n) coded for encultured and situated responses for each category. Of the total responses, 79 were identified as encultured and 51 as situated. In addition, there were six hybrid responses that contained elements of each. For the categories nature of the assessment problem and expectations of assessment, responses coded as encultured were approximately three times more frequent than those coded as situated. Encultured responses expressed expectations of a single, correct answer, did not see the connection to the course material, and conflated ill-defined engineering work with unclear instructions (e.g., “some useful direction as to what models to use may have helped me grasp the context of the assessment better”). Conversely, situated responses generally identified that there were multiple approaches and not necessarily a “correct” answer. These responses identified the need to think creatively and connected the problem to professional settings.

Students were more responsive to the shift in social context, with slightly more responses coded situated as encultured. The situated responses cited a value in learning from their teammates’ perspectives and the way talking

| Nature of assessment problem | Encultured in schooling                                                                                      | Situated in engineering                                                                                     |
|------------------------------|-------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Sample response              | Apparently we all did the midterm exam wrong. Honestly a lot of things that you ask for us to compute are not straight forward and are really confusing. I think I would do much better in the course if you asked for what you are looking for more directly instead of being vague. | The studio felt a bit rushed, but I really enjoyed the content. It made me feel knowledgeable, and it was fulfilling to actually apply my knowledge to a real world situation. |
| Expectations of assessment   | I understand that the studio section was able to add to the grade of the exam, but some useful direction as to what models to use may have helped me grasp the context of the assessment better. | I thought that PART 2 was arguably the most helpful and fairest form of testing that I have had during my time at [this university] and I hope that future exams are similar. Being able to talk out loud what I’m thinking has always been a strong suit of mine, and an individual exam doesn’t allow me to use this skill. |
| Social context/group work    | I did most of the work in my group. It is not fair that we all will get the same grade on it.                 | It is helpful to discuss questions with team mates especially since sometimes I will approach a problem in a way another team mate won’t so this gives me an opportunity to think about a problem from different perspective. |
| Value of experience          | I appreciate the opportunity to make my midterm grade better with the studio assignment                      | Honestly, I should have learned more from this studio. We were fast at making judgements, and though I had some of the conceptual knowledge, I didn’t take enough time to effectively think about the problem. Therefore, I think I am learning to take things more slowly and make better judgements on these engineering problems. I’m bummed that I wasn’t able to go through the problem as detailed as I feel that I should have, but happy that I recognize this now and want to try better next time. |

Table 8

Examples and number of responses (n) coded encultured in schooling and situated in engineering

| Sample response | n  |
|-----------------|----|
| Apparently we all did the midterm exam wrong. Honestly a lot of things that you ask for us to compute are not straight forward and are really confusing. I think I would do much better in the course if you asked for what you are looking for more directly instead of being vague. | 39  |
| The studio felt a bit rushed, but I really enjoyed the content. It made me feel knowledgeable, and it was fulfilling to actually apply my knowledge to a real world situation. | 12  |
| I understand that the studio section was able to add to the grade of the exam, but some useful direction as to what models to use may have helped me grasp the context of the assessment better. | 21  |
| I thought that PART 2 was arguably the most helpful and fairest form of testing that I have had during my time at [this university] and I hope that future exams are similar. Being able to talk out loud what I’m thinking has always been a strong suit of mine, and an individual exam doesn’t allow me to use this skill. | 9   |
| I did most of the work in my group. It is not fair that we all will get the same grade on it | 7   |
| It is helpful to discuss questions with team mates especially since sometimes I will approach a problem in a way another team mate won’t so this gives me an opportunity to think about a problem from different perspective. | 10  |
| I appreciate the opportunity to make my midterm grade better with the studio assignment | 12  |
| Honestly, I should have learned more from this studio. We were fast at making judgements, and though I had some of the conceptual knowledge, I didn’t take enough time to effectively think about the problem. Therefore, I think I am learning to take things more slowly and make better judgements on these engineering problems. I’m bummed that I wasn’t able to go through the problem as detailed as I feel that I should have, but happy that I recognize this now and want to try better next time. | 20  |

5.2 Theoretical conjecture 2: Student assessment experience

We next summarize the main findings from the analysis of student reflection responses to the “Exam Vent” prompt. We observe two points of view in students’ responses, encultured in schooling and situated in engineering. Table 8 presents sample quotations for each category with one column representative of encultured in schooling and the next situated in engineering. Responses coded as encultured focused on school work and the importance of grades, whereas those coded as situated were reflective of doing real engineering work and often connected to engineering identity or understanding. Not all students engaged in the group stage in the spirit of an authentic engineering task.

Table 8 also shows the number of responses (n) coded for encultured and situated responses for each category. Of the total responses, 79 were identified as encultured and 51 as situated. In addition, there were six hybrid responses that contained elements of each. For the categories nature of the assessment problem and expectations of assessment, responses coded as encultured were approximately three times more frequent than those coded as situated. Encultured responses expressed expectations of a single, correct answer, did not see the connection to the course material, and conflated ill-defined engineering work with unclear instructions (e.g., “some useful direction as to what models to use may have helped me grasp the context of the assessment better”). Conversely, situated responses generally identified that there were multiple approaches and not necessarily a “correct” answer. These responses identified the need to think creatively and connected the problem to professional settings.

Students were more responsive to the shift in social context, with slightly more responses coded situated as encultured. The situated responses cited a value in learning from their teammates’ perspectives and the way talking
aloud helped their thinking while the encultured responses pointed more toward a single person driving the solution decisions. Similar views were expressed by students in Cohort 2, who were asked specifically about group work in their Thursday post-class reflection.

In remarking on the value of the assessment experience, approximately twice as many responses aligned with a situated perspective as an encultured perspective. The situated responses cited benefits of realistic engineering work expressing the idea that knowledge is iterative, progressive, and gained by experience. For example, the following responses juxtapose encultured and situated perspectives of reporting an interval estimate:

Not a lot of direction was explicit in the instructions. If we had known that we were expected to make a confidence interval, we would have done so. [Encultured in schooling]

I learned how to apply what we learned to real-life situations. In this studio, my group used confidence intervals to ensure that our reaction rate constant was within a desirable range. [Situated in engineering]

Students coded as situated often connected the value of the second stage to its assessment aim of having them make engineering decisions, as described by the following student:

I learned that analyzing data is a lot more open-ended than I expected. Throughout the studio, my group members and I understood what we needed to do and the general process on how to reach that conclusion. What I did not expect was the various ways to go about it and how multiple ways could be right (some being more right). I thought that determining which method was the best was the most difficult part of the analysis. [Situated in engineering]

6 | DISCUSSION

Guided by a DBR methodology, we present two iterations of analysis of performance and reflection data for engineering students who completed a two-stage midterm exam in a large engineering class required for graduation. The assessment design reported here differs from common manifestations of two-stage exams reported in the literature (Gilley & Clarkston, 2014; Rieger & Heiner, 2014). Rather than having student teams revisit the questions they answered as individuals, we reimagined assessment using an authentic, computer-based task designed to place teams in the role of engineers (Villarroel et al., 2020; Wiggins, 1990).

As illustrated in Figure 4, this study was built upon two high-level conjectures: (i) learning to solve classroom problems does not translate into solving engineering problems in practice (Jonassen et al., 2006) and (ii) authentic, technology-based assessments can allow educators to reimagine assessment (Timmis et al., 2016; Wiggins, 1990). Through embodiment in the tools and materials, task structure, and participant structure of the second exam stage, we explored two design conjectures and two theoretical conjectures, which are discussed in the next sections.

6.1 | Design conjecture 1: Assessment of engineering decisions

The different characteristic solution paths that teams used in responding to the task provide evidence that the task elicits their engineering decision-making in the face of uncertainty (Pickering, 1995; Trevelyan, 2014; Vincenti, 1990). As illustrated by the Sankey diagrams in Figure 6, teams exhibited multiple solution paths based on their choices. Thus, the assessment aligned with important characteristics of engineering practice that we are trying to develop—engineering decision-making. Students identified this aspect as different from their common testing experiences in their reflections.

When faced with an authentic task, many teams struggled to make appropriate engineering decisions such as identifying to use interval estimates, accounting for process variation in making a time recommendation, testing for statistical difference among reactors, or analyzing reactors separately when there was clearly no difference. These results suggest students need more practice operationalizing core course content to make engineering decisions. Such struggles are consistent with situated perspectives of learning. Rather than viewing knowledge as an abstract entity to be “acquired” and then “used,” a situated learning perspective suggests that knowing entails meaningful participation in activities situated within authentic practice (J. Brown et al., 1989; Engle & Conant, 2002; Manz et al., 2020). To the degree that
open-ended problem-solving and decision-making in the face of uncertainty are intrinsic to engineering practice itself, utilizing authentic engineering projects and computer technologies can better prepare students to operationalize knowledge in practical, real-world situations. In addition, by including these practices as a core part of assessment, they likely become more central to engineering instruction throughout the curriculum (Biggs, 1996; James, 2014) and not merely a supporting objective achieved semi-independently by students via internships and undergraduate research. We propose that with regular use, authentic assessments like the one presented here would encourage instructors to shift classroom activities in ways that better prepare students for the engineering profession (Jonassen et al., 2006).

6.2  |  Design conjecture 2: Assessment of procedural accuracy

In addition to assessing a team’s engineering decisions, the computer system affords the ability to evaluate the procedural accuracy for many solution paths by determining if a team’s submitted numerical answers are consistent with the data they were provided given the solution path that they selected. The numerical performance scores were similar to Stage 1, but it was difficult to compare procedural accuracy across teams as some paths inherently led to more challenging calculation procedures. For a subset of cases, we could not verify the team performed their calculations accurately, such as for the teams that estimated a reactor time directly from the raw data. In addition, we identified a few cases where the numerical answers did not match, but where the team pursued a creative solution outside those that we previously considered.

There are challenges in integrating this aspect into an automated scoring system. The more open ended the task, the more possible alternatives that need to be considered. There is potential to make automated scoring more tractable with greater constraints (e.g., specifying the significance level). However, a better understanding is needed in the trade-offs between posing such constraints and the broader pedagogical and epistemological goals of this type of open-ended computer-based authentic assessment.

6.3  |  Theoretical conjecture 1: Student competencies

A Pearson’s correlation analysis suggests that the second stage measures different aspects of engineering knowledge than the first stage and the other individual assessments. Team performance in decision-making for rate constant determination correlated with their decision-making for production time \((r = .47)\) but did not correlate with traditional summative (Final exam, Stage 1) or formative (audience response system) assessments. However, those traditional assessments all correlated with one another. The positive correlation in teams’ decision-making through a design iteration as the technology was improved from Cohort 1 to Cohort 2 adds to the trustworthiness. The method and context of the Stage 2 assessment align more closely with engineering practice than Stage 1.

These correlational results are consistent with Jonassen et al.’s (2006) assertion that professional engineering problems are substantively different from typical in-class problems. While many students demonstrated proficiency with the procedural calculations in traditional assessments, they were not able to operationalize that same content when they needed to make an engineering decision in Stage 2. This finding suggests that knowledge needed in engineering goes beyond procedural competency or even conceptual understanding, but students too need to develop understanding to appropriately apply that knowledge strategically to make decisions in the context of real engineering work (Shavelson et al., 2003). In that vein, it would be useful to investigate how Stage 1 and Stage 2 performances correlate to other professional performance experiences such as in capstone design, undergraduate research, or internships. For many teams, we were able to determine whether they completed their calculations accurately; however, their performance did not correlate with either decision-making or traditional assessment. We believe the lack of correlation results from varying difficulty of calculations for different solution paths, but more research is needed.

6.4  |  Theoretical conjecture 2: Student assessment experience

Importantly, the computer-based second stage provides a holistic assessment (Biggs, 2014) that shifts the messages that students implicitly receive about valued practices in the classroom (Boud, 1990). Consistent with Guzzomi et al. (2017), students expressed varying responses to the shift in assessment practice. On the one hand, some acknowledged how the assessment mirrored the legitimate ways knowledge is used in the profession; for others, however, it violated the norms
of how learning should be measured in school. Like other forms of two-stage exams reported in the literature, students identified using social competencies needed in the professional workplace (Jang, 2016). Specifically, some expressed the value of working in teams where members could bring diverse perspectives and expertise to bear on the complex task (Lotan, 2003; Smith, 1996). Some students embraced the creative and challenging aspects of the situated task and claimed this assessment aligned with their capabilities better than traditional sequestered exams. However, other students struggled with the shift in norms of the authentic assessment, especially around the open-endedness and uncertainty, often in impassioned ways. They associated the uncertainty with a lack of clarity in how the task was presented to them rather than as an inherent aspect of engineering work. Prior school experiences enculture students to expect single correct answers and more obvious connections to the prior work in class (Doyle, 1988; Lampert, 1990; Schoenfeld, 1988). Clearly, this type of assessment disrupts the status quo. As the two-stage exam structure modeled in this study could fit within many large enrollment courses, it provides opportunity to give students regular exposure to more open-ended work. Such experiences can lead to shifts in students’ notions of assessment and of engineering. More research is needed on how such programmatic shifts change students’ perceptions of engineering practice and thereby influence their epistemological commitments.

6.5 Limitations

This study has several limitations. First, it was conducted at one institution and used a single problem. The studio structure of the course studied was particularly amenable to delivery of the second stage reported here, and problem characteristics are known to elicit different responses (Douglas et al., 2012; Jonassen, 2000). While two cohorts showed similar decisions, solution paths, and post-exam reflections, implementation of this two-stage exam structure in other settings with different problems is needed. Second, 59 students from Cohort 1 were excluded as they worked independently, while a technology modification led to only five students being excluded from Cohort 2. Even though both cohorts demonstrated similar decision-making characteristics, this discrepancy leads to sampling concerns when comparing cohorts. Third, while this task was open-ended compared with traditional assessments in engineering science classes, it was quite constrained relative to the messy work of engineering practice (Pickering, 1995; Vincenti, 1990). Fourth, the system reported here is in earlier stages of development, and teams’ decisions were coded manually by researchers. The high reliability, as measured by Fleiss’s kappa, contributes to the study’s trustworthiness. However, our ultimate goal is to automate this type of authentic computer-based assessment and enable widespread delivery in large enrollment engineering science courses. Such tools would assess students’ ability to grapple with uncertainties in engineering work and correspondingly provide instructors a broad understanding of the engineering decisions that teams are making. While automated coding is necessary for widespread implementation, the high reliability obtained in this study combined with ongoing advances in lexical systems (Ha et al., 2011), machine learning (Jescovitch et al., 2020), and artificial intelligence (Johri, 2020; Roll & Wylie, 2016) make this aim appear tenable. Engagement of researchers and developers from these communities is needed. Fifth, while this format shows promise to assess social competencies associated with group work, more effort is needed to identify and refine the processes and practices associated with this aspect of the two-stage exam.

6.6 Contributions

This study offers two major contributions. First, we implemented a technology-based tool and explicated how the tool was used to reimagine assessment. Following the idea that assessment drives the learning (Felder & Brent, 2016; James, 2014; Kahn & O’Rourke, 2005), this study examined an innovative authentic assessment approach for a large engineering science class that placed student teams in the role of engineers doing realistic work. The focus of the assessment shifted from conceptual understanding and procedural accuracy to the measure of teams’ ability to manage uncertainty to make decisions. A Pearson’s correlation analysis suggests that the authentic assessment in the second stage measured different competencies than the traditional classroom assessments, supporting the high-level conjecture that learning to solve classroom problems does not translate into solving engineering problems in practice. While some students embraced the creative and challenging aspects of the situated task, other students struggled with the open-endedness and uncertainty, often in impassioned ways. In shifting assessment practices, instructors need to address explicitly classroom norms and student expectations.
Second, we apply the methodological approach of DBR to a learning system in engineering education. DBR is particularly appropriate for instructional innovations pointed toward professional formation of engineers (Dasgupta, 2019; Diefes-Dux et al., 2010; Gomez & Sviha, 2019; Minichiello & Caldwell, 2021; Newstetter, 2005; Weber et al., 2014) and where learning engineering is mediated by technology tools (Friedrichsen et al., 2017; Minichiello & Caldwell, 2021). Following Sandoval (2014), we used conjecture mapping to explore simultaneously design conjectures and theoretical conjectures through two design iterations. Our approach aligns with DBR criteria for computer-based tools posed by Jeong et al. (2014) to be set in a specific context, grounded in theory, and be part of a larger DBR program. However, there are important differences between this study and DBR studies reported in the literature. Commonly, DBR studies addressing classroom learning focus on shifts in student reasoning and argumentation (e.g., Reimann, 2011; Sandoval, 2014). This orientation has appropriately led to the examination of the discursive practices between students with one another and with the instructor. We agree that these characteristics of classroom learning are important in the formation of engineers, and other studies from our larger initiative focus on discursive processes during classroom learning (Hirshfield & Koretsky, 2021; Koretsky et al., 2021). However, equally important, yet uncommon in DBR studies, are the assessment processes that drive the learning. At the same time, we see merit in examining discursive processes during the team-based assessments described here as a fruitful area for future research.

ACKNOWLEDGMENTS
The authors gratefully acknowledge the support provided by the National Science Foundation through Grant EEC 1519467. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation.

ORCID
Milo D. Koretsky https://orcid.org/0000-0002-6887-4527

ENDNOTE
1 The term “engineering science” course is chosen to be consistent with the historical language in US ABET accreditation guidelines.

REFERENCES
Bakker, A. (2018). Design research in education: A practical guide for early career researchers. Routledge. https://doi.org/10.4324/9780203701010
Barab, S., Dodge, T., Tuzun, H., Job-Sluder, K., Jackson, C., Arici, A., Job-Sluder, L., Carteaux, R., Jr., Gilbertson, J., & Heiselt, C. (2007). The Quest Atlantis Project: A socially responsive play space for learning. In B. E. Shelton & D. Wiley (Eds.), The educational design and use of simulation computer games (pp. 159–186). Sense Publishers. https://doi.org/10.1163/9789087903121_011
Barab, S., & Duffy, T. (2000). From practice fields to communities of practice. Theoretical Foundations of Learning Environments, 1(1), 25–55.
Bearman, M., Dawson, P., Ajjawi, R., Tai, J., & Boud, D. (2020). Re-imagining university assessment in a digital world. Springer. https://doi.org/10.1007/978-3-030-41956-1
Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. The Annals of the American Academy of Political and Social Science, 683(1), 217–232. https://doi.org/10.1177/0002716219846850
Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. http://doi.org/10.1007/BF00138871
Biggs, J. (2014). Constructive alignment in university teaching. HERDSA Review of Higher Education, 1, 5–22.
Boaler, J., & Greeneo, J. G. (2000). Identity, agency, and knowing in mathematics worlds. In J. Boaler (Ed.), Multiple perspectives on mathematics teaching and learning (pp. 171–200). Ablex.
Boud, D. (1990). Assessment and the promotion of academic values. Studies in Higher Education, 15(1), 101–111. http://doi.org/10.1080/03075079012331377621
Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. CBE—Life Sciences Education, 14(2), es4. http://doi.org/10.1187/cbe.14-11-0208
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. http://doi.org/10.1111/1478-8787.00630
Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. Journal of the Learning Sciences, 2(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2
Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42. https://doi.org/10.3102/0031389X018001032
Bucciarelli, L. L. (2002). Between thought and object in engineering design. Design Studies, 23(3), 219–231. https://doi.org/10.1016/S0142-694X(01)00035-7
Bucciarelli, L. L. (2003). Engineering philosophy. Delft University Press.
Manz, E., Lehrer, R., & Schauble, L. (2020). Rethinking the classroom science investigation. *Journal of Research in Science Teaching, 57*(7), 1148–1174. https://doi.org/10.1002/tea.21625

Manz, E., & Suárez, E. (2018). Supporting teachers to negotiate uncertainty for science, students, and teaching. *Science Education, 102*(4), 771–795. https://doi.org/10.1002/sce.21343

Mazur, E. (1997). *Peer instruction*. Prentice-Hall.

Michor, E., & Koretsky, M. (2020). Students’ approaches to studying through a situative lens. *Studies in Engineering Education, I*(1), 38–57. http://doi.org/10.21061/see.3

Minichiello, A., & Caldwell, L. (2021). A narrative review of design-based research in engineering education: Opportunities and challenges. *Studies in Engineering Education, I*(2), 31–54. http://doi.org/10.21061/see.15

Mitkov, R., Le An, H., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering, 12*(2), 177–194. https://doi.org/10.1017/S1351324906004177

Murray, J. K., Studer, J. A., Daly, S. R., McKilligan, S., & Seifert, C. M. (2019). Design by taking perspectives: How engineers explore problems. *Journal of Engineering Education, 108*(2), 248–275. https://doi.org/10.1002/jee.20263

Newstetter, W. C. (2005). Designing cognitive apprenticeships for biomedical engineering. *Journal of Engineering Education, 94*(2), 207–213. https://doi.org/10.1002/j.2168-9830.2005.tb00841.x

Pellegrino, J. W., & Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. https://doi.org/10.17226/10019

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for educational and social change: Linking scholarship, policy, and practice (pp. 37–50). Springer.

Rieger, G. W., & Heiner, C. E. (2014). Examinations that support collaborative learning: The students’ perspective. *Journal of College Science Teaching, 43*(4), 41–47. http://doi.org/10.2505/4/jcst14_043_04_41

Riessman, C. K. (2008). *Analyzing qualitative data: Handbooks and software*. Sage Publications.

Rikken, A. K. (2001). Rethinking the classroom science investigation. *Science Education, 85*(4), 311–326. http://doi.org/10.1002/sce.21343

Smith, K. A. (1996). Cooperative learning: Making “group work” work. In T. E. Sutherland & C. C. Bonwell (Eds.), *Using active learning in college classes: A range of options for faculty* (pp. 71–82). Jossey-Bass.

Steams, S. A. (1996). Collaborative exams as learning tools. *College Teaching, 44*(3), 111–112. https://doi.org/10.1080/87567555.1996.9925564

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of ‘well-taught’ mathematics courses. *Educational Psychologist, 23*(2), 145–166. https://doi.org/10.1207/s15326985ep2302_5

Sud, P., West, M., & Zilles, C. (2019). *Educational games to support caring and compassion among youth: A narrative review*. Sage Publications.

Stevens, S. A. (1996). *Collaborative exams as learning tools*. Jossey-Bass.
Our new OrangeCandy product line needs to go into volume production. For the process, we need a source for glucose ($C_6H_{12}O_6$) and fructose ($C_6H_{12}O_6$). We will produce these sugars through a hydrolysis reaction using sucrose ($C_{12}H_{22}O_{11}$).
O_{11}) as a reactant in aqueous solution (0.5 M HCl) with our proprietary RateEnhancer additive that is believed to catalyze the reaction.

The hydrolysis reaction is monitored by a polarimeter. In this technique, the angle of plane-polarized laser light is measured as it is passed through the solution. The change in angle can be related to the concentration of sucrose in the solution.

The biochemists from the consulting firm we hired report that the kinetics for this irreversible first-order reaction can be described by the following ordinary differential equation:

\[
\frac{dC_S}{dt} = -kC_S,
\]

where \(C_S\) is the sucrose concentration in (mol/m\(^3\)), \(k\) is the first-order reaction rate constant in (h\(^{-1}\)), and \(t\) is time in (h). However, they are not able to provide us a value for the rate constant, \(k\), as we do not want to provide them access to our proprietary additive.

The quality we provide to our customers is of utmost importance at Beaver Dam Sweet Treats. The process design team reports that it is critical that at least 70% of the initial sucrose has reacted to make the final product acceptable. Conversion of less than 70% requires reprocessing the entire batch. Due to production bottlenecks, we also need to run the process for as short a time as possible. Due to process flow requirements, our two batch reactors need to use the same process time.

Please determine the rate constant and use it to make a process recommendation that you are confident will reach the 70% conversion requirement. I suggest you do the following analysis prior to experiments.

1. First, the equation above must be solved to get a relationship between concentration and time. Please do this on your whiteboard.
2. Second, concentration versus time data are needed. It is helpful to draw a rough schematic of what the experimental equipment would look like. Please do this on your whiteboard and have it approved by one of your supervisors.

Since each run takes many hours, each team member will only have two experimental runs to collect data (one run on each batch reactor), but our ConceptWarehouse software will allow you to see the data from your teammates as well. Before you do your runs, you must receive approval from a supervisor.

To do a run in the batch reactor, please log onto the Concept Warehouse. It is very important that you put in your correct section and team number to see your teammates’ data and get credit for your run.

Using the cumulative data set collected by your group (and only your group), please:

1. How do you suggest to report the reaction rate constant, \(k\), for the company databank? Please include numbers.
2. Recommend how long you think the operators on the production floor should run the batch reactors to get at least 70% conversion. Suggest a single process time.
3. Justify the values you provide and your recommendation, supporting your analysis with appropriate data.
4. As we change or modify our product line, the sugar specifications change. Develop a model that predicts the run time needed for any conversion the production supervisor may want in the future and report it.
As a **team**, please provide all software files you generated for analysis to Gradescope and as an **individual** report values for $k$, time, justification, and model to the Concept Warehouse.

**APPENDIX B: STAGE 2—RUBRIC PROVIDED TO STUDENTS DURING THE SECOND STAGE**

Team: _____________________________

Guidelines:

- Come to a team consensus on the recommendations
- Do your scratch work on the whiteboard; everybody should participate.
- The entire team must have the opportunity to contribute and reach consensus.
- You will be graded on our assessment of the degree you were working together as a team and your ability to come up with a reasonable answer.
- Your team will turn in supporting work in Gradescope and each individual will provide answers in the Concept Warehouse.

| Problem                      | Score | Possible |
|------------------------------|-------|----------|
| Team collaboration and engagement | 6     |          |
| $k$                          | 2     |          |
| Production time              | 3     |          |
| Justification                | 2     |          |
| Model                        | 1     |          |
| **Total**                    | **14**|          |

(6 pts) Instructor evaluation of teaming process and group engagement

| Teaming            | Rubric                                                                 | Score |
|--------------------|------------------------------------------------------------------------|-------|
|                    | All team members are actively participating, listening to others, and sharing their ideas | 3     |
|                    | The work is mostly distributed with all having a chance to participate. | 1.5   |
|                    | The work and approach are dominated by one individual                 | 0     |

| Engagement         | Rubric                                                                 | Score |
|--------------------|------------------------------------------------------------------------|-------|
|                    | The team is actively collaborating and working toward a solution the entire time. | 3     |
|                    | Not all team members are active; or slight periods of disengagement    | 1.5   |
|                    | Substantial periods of disengagement                                   | 0     |
APPENDIX C: SAMPLE ITEMS FROM THE CLASSROOM ASSESSMENTS

Figures C1–C4 provide illustrative items from classroom assessments including the first stage of the midterm exam and the audience response system. For each case, an example focused on conceptual understanding and an example focused on procedural accuracy is shown.

**FIGURE C1** Sample item from Stage 1 (individual) portion of the two-stage exam focused on conceptual understanding [Color figure can be viewed at wileyonlinelibrary.com]

\[
2. \quad (16 \text{ pts}) \text{ The means of samples consisting of 9 measurements have the distribution shown below.}
\]

A. Describe the shape of this distribution.
B. What is the area under the curve? \[\text{Area} = \underline{\quad} \]
C. As accurately as you can, draw the population distribution that corresponds to the distribution above. \textit{Label the important features.}
D. Report values for the following:
\[
\begin{array}{cc}
\mu & \underline{\quad} \\
\sigma & \underline{\quad} \\
\mu_{\overline{x}} & \underline{\quad} \\
\sigma_{\overline{x}} & \underline{\quad}
\end{array}
\]

**FIGURE C2** Sample item from Stage 1 (individual) portion of the two-stage exam focused on procedural accuracy

5. \( (30 \text{ pts}) \text{ You have just invented a process for making a high valued pharmaceutical. You want to estimate how much you can produce so that you can determine appropriate licensing fees. In 10 consecutive days, you get the following amount of product (in kg):} \]

\[\{15, 18, 41, 27, 16, 29, 30, 32, 25, 18\} \]

For this data set, \( s = 8.36 \)
A. To 90% confidence, how much product does your process produce?
B. To 90% confidence, what is the standard deviation of the process?
C. Estimate how many days you would need to operate to reduce your interval (at 90%) to \( \pm 1 \text{ kg.} \)
A large college class has 900 students, broken down into section meetings with 30 students each. On the final exam, scores followed a normal distribution with an average of 65 and a standard deviation of 15.

If we consider a section of 30 students as a random sample from this population, will the probability that the average for the entire section is between 50 and 80 be (greater, the same, lower) as what you calculated in the previous question? (You do not need to perform any additional calculations.)

- The probability will be greater
- The probability will be the same
- The probability will be less

Please explain your answer in the box below:

Please rate how confident you are with your answer.

- Substantially unsure
- Moderately unsure
- Moderately confident
- Substantially confident

Submit

**Figure C3** Sample item from in-class audience response system focused on conceptual understanding [Color figure can be viewed at wileyonlinelibrary.com]

You are an engineer in charge of a chemical mechanical polishing tool in a semiconductor fab. A new product is coming through the pipeline and you need to estimate the mean polish time (μ). Based on previous, similar products that you have run on your tool, you believe that the population of polish times can be approximated by a normal distribution.

Since this product is still in testing, you are only able to measure 6 wafers and get the following polish times (s):

- 33.9
- 35.9
- 43.9
- 41.9
- 31.6
- 42.6

Using the sample data collected above, determine the 95% confidence interval of the population mean of this new product.

- [34.8s - 41.8s]
- [34.2s - 42.4s]
- [34.1s - 42.6s]
- [33.8s - 48.7s]
- [33.2s - 43.5s]
- [32.9s - 43.7s]

Submit

**Figure C4** Sample item from in-class audience response system focused on procedural accuracy [Color figure can be viewed at wileyonlinelibrary.com]