Computers and Education Open 10 (2026) 100379

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Computer-based authentic assessment in engineering: a case of group practice

Milo D. Koretsky [a,*], Irene Pantekidis [b]

[a] Department of chemical and biological engineering, Department of education, Tufts University, 574 Boston Ave, 102B, Medford, MA 02155-5801, USA
[b] Department of child study and human development, Department of biology, Tufts University, Medford, MA 02155-5801, USA

ARTICLE INFO

Keywords:
Authentic assessment
Computer-based assessment
Engineering
Figured worlds
Interaction analysis

ABSTRACT

Traditional assessment practices that focus on lower-level cognitive skills no longer align with the knowledge and social practices students will need as professionals in the era of generative artificial intelligence. In this qualitative study, we analyzed students' approaches and participation in a computer-based group exam for second-year university engineering students in the US. Students were positioned as an engineering group in a hypothetical candy factory collecting reactor data within a Unity 3D computer simulation, representing the industrial setting. Four groups were video recorded as they completed the two-hour exam. Videos and transcripts were analyzed via interaction analysis. In addition, their exam answers and other course assessment scores were collected as secondary data. More successful groups in this study tended to share material tools on their Zoom screens and toggle between engineering decisions, while the least successful group rarely shared material tools and took a sequential approach to decision-making. We observed both single and dual leadership modes, with most students clearly participating in ways to move the solution forward. However, some students did not appear to contribute substantially, raising concerns about using such exams for credentialing purposes. Taking a theoretical lens of group practice within figured worlds, we observed that the individual students became entangled in the group's approach implying that, in addition to the individual, the group should be a fundamental unit of analysis for engineering assessments. In the cases studied, the computer-based tool provided an authentic task that allowed them to demonstrate such group practices.

1. Introduction

Assessment is central to STEM (science, technology, engineering, and mathematics) education, functioning both as a means of gathering and interpreting information to guide teaching and learning and as a tool for credentialing STEM graduates [1,2]. Computer tools have become increasingly used to support assessment processes [3]. These tools range from automating scoring of traditional assessments often with large question banks [4,5] to using student-computer interaction data to provide new sources of assessment information [6,7]. However, many school assessments place students in simplified and unrealistic situations [8–10]. Students are often expected to work alone, use limited resources, and solve problems that have one correct answer and a clear solution path [11–13]; they are often judged by how fast they can complete these assessments [14]. Instructors design assessments this way because it's more practical, especially in large classes [15]. They want the grading to be manageable and the evaluations to feel fair [16, 11]. Additionally, recent advances in generative artificial intelligence (AI) have exacerbated problems with these traditional school assessments. Systems like ChatGPT can calculate correct answers to constrained problems well [17–19], which brings questions of academic integrity of traditional assessments into play [20–24].

These issues necessitate the need to radically rethink the purpose, design, and implementation of assessment practices. One opportunity is to leverage a computer-based tool to provide students with an *authentic assessment* experience [25–28], using the tool to both provide different cases among students within the same class and to assess student performance [29–32]. An authentic computer-based assessment task should require students to use the kind of reasoning and problem-solving skills they will use in their careers. In order to relate to how students will apply their knowledge in the future and encourage self-reflection, the task should be realistic, partly open-ended (so students must define the problem themselves), and allow for multiple possible solutions [27]. Success on these assessments then requires demonstrating deep

* Corresponding author.
E-mail address: milo.koretsky@tufts.edu (M.D. Koretsky).

https://doi.org/10.1016/j.caeo.2026.100379
Received 12 September 2025; Received in revised form 10 May 2026; Accepted 10 May 2026
Available online 11 May 2026

understanding of the subject, thinking critically, and solving complex problems [33,34].

In the case of engineering education, authentic assessment involves creating a task that reflects the real work engineers do. Practicing engineers are judged by how well they can deliver a solution that meets a set of goals and constraints [35]. Their work involves uncertainty and incomplete information, which means there can be many ways to solve a problem [36,37,35]. Ideally, the assessment would also reflect the social practices of engineering [38,39], where students work in groups, allowing the group members' different perspectives to potentially add value to the work [40,41].

Previously, we reported such an innovative computer-based authentic assessment for a large second-year university engineering class [29]. Based on quantitative analysis of 117 groups (381 students), we found that this group exam shifted the focus from procedural and conceptual knowledge to engineering decision making, better aligning the assessment with engineering practice. While that study elucidated the different decisions that the groups made, it did not reveal the *processes* by which they made those decisions. In the study presented in this article, we use video data of four groups together with an interaction analysis to reveal how students interacted with the task and with their group members to make decisions. This understanding is critical to the design of effective computer-based authentic assessments.

2. Background

2.1. Group practice

In professional practice, a single individual rarely completes an engineering project [39]. Instead, the work is collaborative and sociotechnical. Sociotechnical refers to how engineered systems integrate into communities and society, hopefully with positive outcomes but sometimes with detrimental consequences [42–44]. Sociotechnical also refers to how the work itself is organized. The relative status and authority of members shape the social processes that ultimately guide the group's solution path [45–47]. Collaboration in engineering is supported when those social practices draw upon the distributed expertise and varied perspectives of the group members.

Drawing on social practice theory, Medina and Stahl [48] highlight group practice, or thinking at the collective level, as distinct from individual practice. Group practice involves the socially negotiated ways people coordinate their work [49,50]. For example, a group may coalesce around a single leader or distribute authority more evenly to support engagement with one another's ideas [51]. Thus, meaningful work occurs through "intertwined levels of individual, small-group, and community processes" ([48], p. 201). These patterns often develop tacitly, outside members' awareness.

As with professional engineers, students' success on authentic classroom tasks depends heavily on their group's social practices. Yet, few studies have examined how these unfold, especially among undergraduates. Barron [52] found that stronger groups actively engaged with members' competing ideas, while weaker groups ignored one another's ideas. Such group practice predicted high-quality solutions more strongly than members' prior achievement or the number of ideas generated. Menekse et al [49] reported similar results for engineering students. Jordan and McDaniel Jr [53] showed that social practices also help members manage uncertainty and move forward despite incomplete knowledge. Less productive social patterns have been reported as well. Kittleson and Southerland [54] describe "divide and conquer" strategies that limit interaction and shared knowledge building. Students' prior schooling also matters; those used to single-answer problems may react with frustration or controlling behaviors when faced with open-ended tasks [47]. Such dynamics can heighten gender inequities in the classroom [55,56,47].

From an assessment perspective, group practice must be evaluated in a group context. However, group work often doesn't allow students to be assessed individually, and students can become frustrated that these tasks aren't designed to fairly reflect each person's contribution [57,58]. This frustration results from common problems students perceive in group work, such as unequal participation, where some members contribute less (free riding or social loafing) or others feel taken advantage of [59]. Research shows that students are often concerned that everyone in the group gets the same grade, even when the amount of effort varies [60,61].

2.2. Figured worlds

We use *figured worlds* [62,63] to frame the analysis of group practice during the authentic assessment. Figured worlds are socially and culturally shaped environments where people recognize certain roles, give meaning to actions, and value specific outcomes [63]. These are shared "realities" where people act as if certain things are true or important. For example, the figured world of engineering school ("School World") includes common roles like students, instructors, and advisors [64,65]. In this world, learning usually means getting the right answers by following given methods, with high grades seen as proof of success. People act as if this process prepares students to become engineers. In contrast, the figured world of professional engineering ("Engineering World") involves roles like clients, coworkers, and managers, and the work includes collaborating on teams, dealing with deadlines and budgets, and creating designs that meet real-world needs. In this world, the goal isn't just getting the "right" answer; rather activity focuses on solving problems within real constraints.

In each world, actions are judged by what's considered important within that setting. For example, trying out multiple approaches to solve a problem might be valued in Engineering World but not in School World, where one right answer is expected. Figured worlds are shaped and maintained through participation, using tools and artifacts that support activities [66,67]. In workplaces, this might be a memo or contract; in schools, it's often worksheets or handouts. These material tools structure how people participate. In School World, worksheets reflect the instructor's authority and guide students toward producing the "correct" answers that lead to good grades. These tools carry meaning because of their role within the broader figured world of schooling.

2.3. Research questions

We select the framing of figured worlds to acknowledge the duality of this group exam as being a School World assessment that seeks to promote authentic Engineering World student activity. We seek to understand how four groups of university engineering students interact with material resources and with each other to make decisions. Ultimately our analysis focuses on both the assessment part (How do members of a group participate? Is this fair relative to the traditional exam where individuals are held accountable?) and the authentic part (To what degree do students take up roles of engineers solving a realistic task? To what degree do they approach the exam as a school task?)

To guide this inquiry, we pose the following specific research questions:

1. What solution paths did groups take as they completed the authentic engineering exam? What material and knowledge resources did they utilize in decision-making?
2. What social roles did groups take on? How did these roles influence participation?
3. How did the students' framing and group dynamics influence their collaborative decision-making?

3. Task design

For this authentic assessment, students were positioned as an engineering group optimizing a production process in a hypothetical candy factory. The assessment task embeds process variation, measurement variation, multiple possible solution paths, and opportunities for engineering judgment, all of which make it more authentic than conventional exams. Specifically, students were tasked with (1) determining the rate constant, k, for a sugar reaction from a set of experimental data and (2) using the resulting values to recommend the run time for two hydrolysis reactors. They worked in groups of 3 to 5, each collecting data within a simple Unity 3D computer simulation, representing the industrial setting (see Fig. 1). Students engaged in the simulation in “first-person” mode, using virtual equipment to operate reactors (Fig. 1a) and take measurements (Fig. 1b). To support the industrial scenario, the assignment is framed as a memo from the Vice President of Engineering. Groups are provided an “equipment manual” with instructions for running the reactors and collecting data.

The simulation runs in regular web browsers using WebGL software. It uses a kinetic model to generate reactor data, with added random noise. The first principles model lets groups practice scientific reasoning, while the added noise allows for statistical reasoning. To be successful, groups need to toggle between these types of reasoning. Each group has access to two reactors, and the simulation gives each group member slightly different results based on the added noise. Each group gets unique reactor performance data characterized by different values of the rate constant, k. Sometimes, the two reactors for a given group have the same reaction rate constant and should behave similarly except for the noise. Other times, the reactor kinetics are different by design. The group needs to account for two sources of variation in the data: (1) variation between the two reactors and (2) variation between each member collecting data on a given reactor. Groups can download a spreadsheet that provides all members’ time-concentration data from the simulation. The students’ parameters, data, and responses are saved in a central database.

The groups need to make several decisions as they complete the task [29].

1. To analyze the data, they need to choose the *Mathematical Form* for the rate expression that describes the concentration-time relationship. The derived form of the rate equation is exponential, but a logarithmic transformation allows linear regression methods.
2. The groups need to develop a *Solution Strategy* for organizing the data collected from two reactors and multiple group members to perform their analysis.
3. For the *Determination of Rate Constant*, they could choose to provide a point estimate (single value) or an interval estimate (confidence interval).
4. In the *Time Calculation* for production, they could account for process variation in selecting a process time for manufacturing or simply use the point value of the rate constant.
5. Finally, they could perform a *Statistical Analysis* to determine if the two reactors’ rate constants were statistically different, they could eyeball the difference, or they could not explicitly address reactor differences.

Each group submitted a video of their work on Zoom which was evaluated according to a rubric (Appendix A). The video was spot checked for collaboration and engagement, paying particular attention to the period after they collected data. The decisions and justification were evaluated based on their submission through the student interface.

4. Methods

4.1. Study design

We used interaction analysis [68,69] to characterize how students were engaging with the computer-based task and with one another to make engineering decisions. Interaction analysis is a qualitative research method used to study how people interact with each other, objects, and tools in natural settings, typically using video recordings as primary data. Building upon traditions in sociocultural theory, ethnomethodology, and conversation analysis, interaction analysis centers on closely examining moment-to-moment social interactions to understand how meaning, knowledge, and action are constructed collaboratively over time. Through the methodological lens of interaction analysis, we viewed the use of material tools as human-task mediated influences and the emergent social roles as human-human mediated influences. We focused on how students negotiate design decisions, how tools mediated communication, and how leadership and other roles emerged through talk and gesture. In addition, we used exam submissions and other course performance data to support our interpretations and theorizing.

An overview of the methodological approach is shown in Fig. 2 and described next.

4.2. Participants and setting

The computer-based group exam was conducted in an engineering course that is required for students majoring in chemical, biological, or

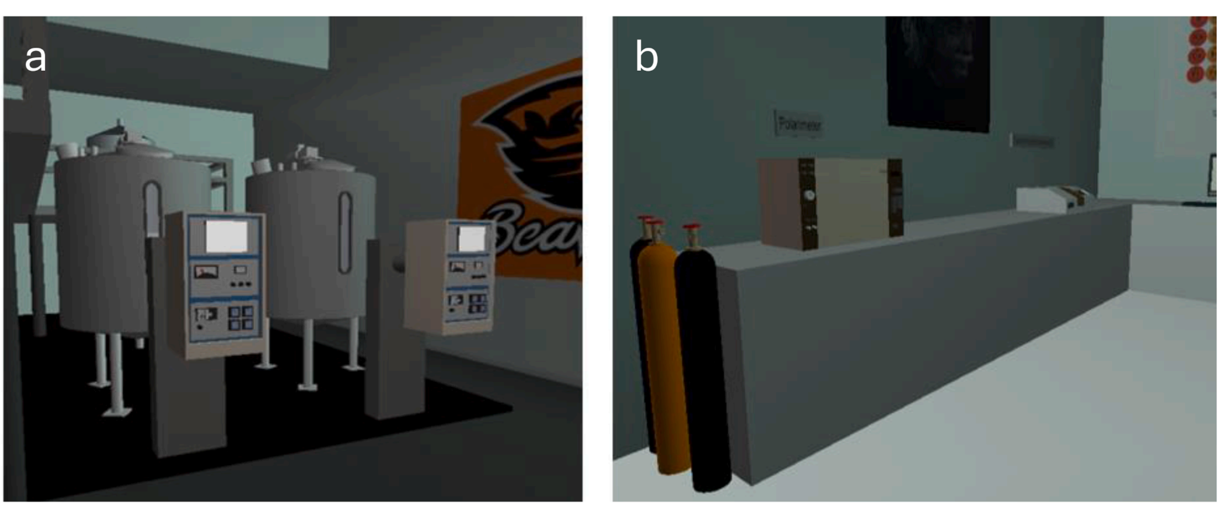

Fig. 1. Screenshots of the 3D laboratory simulation show the two working areas. (a) the reactor bay, which has two hydrolysis reactors and control panels and (b) the analytical lab, which includes an automated polarimeter and other analysis tools.

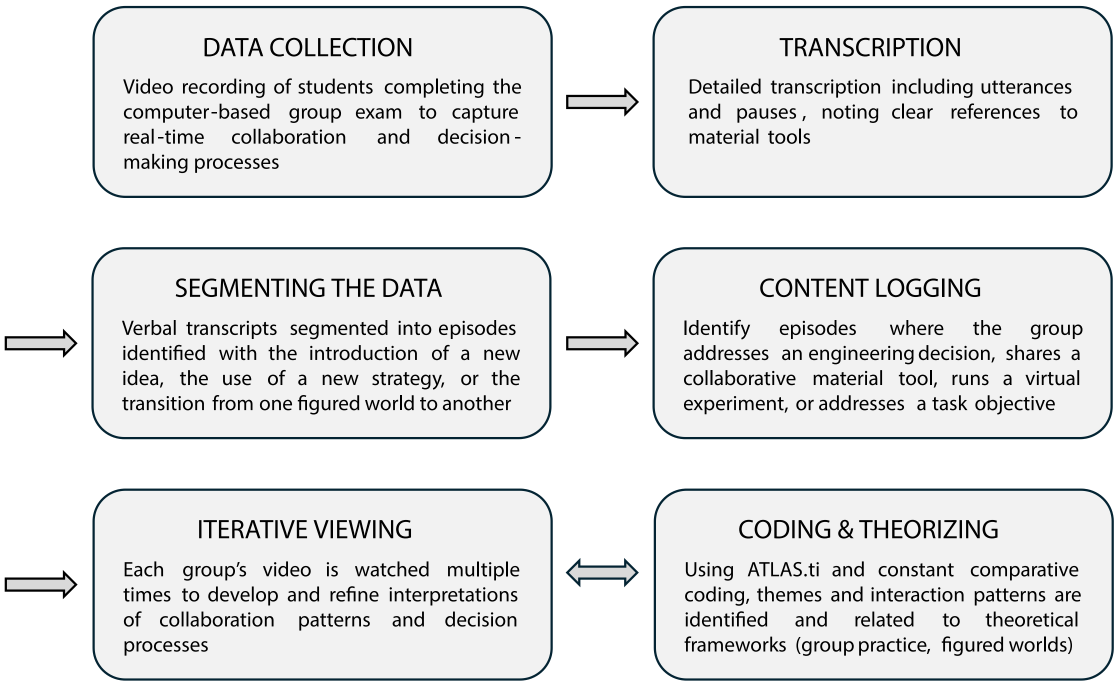

Fig. 2. Steps in interaction analysis of the computer-based group exam.

environmental engineering at a public university in the Pacific Northwestern United States. The enrolled students were mostly second-year students, but some students were in later years of study. This exam was conducted in the sixth week of the term after five weeks of lecture content delivery, related readings and comprehension assessments, and studio sections.

The course was delivered in Spring 2020 in emergency remote mode due to the Covid-19 pandemic, and the instructor viewed the computer-based exam as an asset under these circumstances. Due to the uncertainties in engagement during the pandemic, the instructor increased the team sizes by one from the previous study - from three to four students in the prior in person delivery to four to five students in remote mode [29]. One of the target groups had a student withdraw from the course resulting in the sizes of the four groups studied. They were the same groups that had been used in the prior five weeks during studio sessions. The day before the computer-based group exam, students individually completed a 20-question multiple-choice exam where questions were drawn from a large question bank. The class was synchronously delivered on Zoom and was taught interactively with frequent use of an audience response system (ARS) [70]. More details of the course are presented in [71].

Students used laptops to access the exam materials including the simulation to submit their answers. The exam was also open-book and open-internet, with the limitation that humans other than group members or the instructional team could not be consulted. All groups were video recorded on Zoom. To increase the transferability of the interaction analyses, the four groups were selected to provide a distribution in both exam performance and group demographics. The research was approved by the university IRB and all participants provided informed consent. Pseudonyms of the participants including self-reported demographic information are shown in Table 1.

Table 1
Group participants including demographics and individual performances.

Group	Pseudonym	Gender	Ethnicity	Midterm (Individual) out of 20 possible	Audience Response System Questions Correct (%)[a]
Group A	Aaron (S1)	Man	White	18	64.3
	Alex (S2)	Man	White	18	68.8
	Austin (S3)	Man	White	12	81.2
	Angela (S4)	Woman	White	17	75.0
Group B	Beto (S1)	Man	LatinX	18	64.3
	Baily (S2)	Woman	Chinese American	16	62.5
	Barbara (S3)	Woman	White	17	62.5
	Bonnie (S4)	Woman	White	16	75.0
Group C	Caleb (S1)	Man	White	19	56.3
	Clara (S2)	Woman	White	17	31.3
	Conner (S3)	Man	White	15	50.0
	Chenxi (S4)	Woman	Chinese	16	62.5
	Cody (S5)	Man	White	16	53.3
Group D	Dahlia (S1)	Woman	White	17	28.6
	Deetya (S2)	Woman	Indian American	11	62.5
	Dakota (S3)	Woman	White	11	43.8

[a] . The percentage correct for ARS questions over the entire term is reported.

4.3. Data sources

Video recordings were transcribed verbatim including utterances and notable pauses and use of material tools. We use the time elapsed into the exam when reporting utterances. For secondary data, we collected the groups' exam submissions, each student's individual midterm score from the previous day (out of 20 total), and the percent correct for all ARS questions over the term. The latter two are shown for each student in Table 1.

4.4. Positionality and trustworthiness

The first author, an able-bodied, cisgender, straight, White man, was the lead instructor for the course in which the group exam was delivered. He leads a research program that uses technology tools to develop learning systems that challenge students to integrate and extend the knowledge they have learned in the core curriculum to the more complex, realistic problems they will encounter as professionals. The system reported here, and analysis thereof, builds on his prior work on computer-based projects [55,72–75] and class activities [76–78] but in novel ways that seek to position computer-based tools as supports for novel and authentic assessment needed in the 21st Century.

The second author, an able-bodied, cisgender, straight, White woman, has BS degrees in Child Study and Human Development, Biology, and Spanish Cultural Studies. She approached this research from the perspective of both a student and a STEM teaching assistant, where she worked regularly with groups of students as they collaboratively approached questions in class. She has completed coursework on how to effectively teach and academically guide others, which encouraged her to think deeply about how to create equitable STEM education environments, promote collaborative learning amongst students, and help prepare students for both higher-level courses and continued post-graduate success. Many of these concepts informed her perspectives, ideas, and motivations for this research.

Rather than an independent interpretation of coder reliability, the authors worked throughout data collection and analysis to reduce bias, ensure high-quality research, and make sure our coding system could be applied in other settings [79,80]. To achieve this, the two authors met regularly over a three-year timespan to review both the video recordings and the verbatim transcripts. These meetings included collaborative coding in ATLAS.ti as well as thorough discussions of their interpretations of each group's social dynamics and decision-making processes, as summarized in analytic memos. The conversational and qualitative methodological approach we chose allowed for continuous exchange of the two authors' perspectives.

We supported analytic consistency by collaboratively developing code definitions, having regular meetings to compare interpretations across authors, revisiting earlier coded data when categories changed, and using analytic memos to document major decisions as they occurred during the coding process. Disagreements that occurred in regard to transcript interpretation were usually resolved by viewing the original video and accounting for tone and body language. In only a few cases, we noted two potential interpretations based on each author's perspective. The appendices provide the exam rubric, the codebook, and a coded transcript excerpt to make the analytic approach more transparent and accessible.

While the authors tried to consider the unique group and power dynamics that influence academic collaborative work, they recognize that their collective experiences inform and affect how they analyze students' social interactions. As a female student, the second author found that many of her perspectives on gender power dynamics were informed by her own experiences in STEM education. However, as privileged White researchers, we acknowledge the limitations of our experiences and the effects these limitations have on the subjectively informed interpretations of the groups completing the described group exam.

4.5. Data analysis

Coding was conducted using ATLAS.ti (version 9) software and occurred in several phases (See Fig. 2). First, the transcripts were parsed into episodes, thematic units of discourse that have a central theme and a relatively clear beginning and end [81–83]. Following Gilbuena et al [81], episode boundaries were delineated by the introduction of a new idea, the use of a new strategy, or the transition from one figured world to another.

Next, we performed content logging of each group's activity by examining the content of the episode. We logged four aspects of activity that were present: the type of decision the group was addressing (Mathematical Form, Solution Strategy, Rate Constant, Time Calculation, Reactor Difference, Confidence Interval); any resource that was shared on the Zoom screen (White Board, Excel Sheet, Virtual Simulation, Web Resource, GoogleDoc, Assignment); the group's virtual experimentation; and any reference to project deliverables (rate constant, production time, approach).

We then engaged in two rounds of open and axial coding of video and transcript data using constant comparative coding [84]. Broad categories included decision, social, and figured world (see Appendix B for the code book). A combination of emergent and a priori codes were used. The subcodes identified in the Social category (e.g., Confirming question, Negotiation, New idea introduced …) emerged from analysis of the group's social interactions. In contrast, Figured Worlds subcodes were predetermined based on our theoretical framework (see Section 2.2). For the Decisions subcodes, we began with the five decisions identified in the prior study [29] but supplemented those subcodes with two new decision types that emerged from the current analysis (Significant figures, Solving equation). A sample coded transcript showing one code from each major category is provided in Appendix C. The sample was selected since it was brief but contained a subcode from each of the major code categories (Social, Figured Worlds, Decision).

The first round of coding focused on the analysis of group dynamics and interactions, as well as figured worlds. These included individual and group moves such as introducing a new idea, collaboration across group members, and regulatory moves. The second round of coding centered on an analysis of each group's decision-making processes and how this connected to the group dynamics/interactions identified in the first round. Each decision identified in the content logging was individually analyzed. We consistently wrote analytic memos with our interpretations and perspectives on the group's human-task and human-human interactions. This process occurred collaboratively for Groups A and D, and independently by the first author for Groups B and C. To support and elaborate on our descriptions, the excerpts reported in Section 5.3 were discussed line-by-line by the first author's research group in two group meetings after the initial draft of these cases was written. The research group consisted of one postdoctoral scholar, six PhD students, and two undergraduate students. Finally, after the preliminary draft of the Findings was written, the first author rewatched each video, looking for confirming or disconfirming cases.

5. Results

5.1. Decision-Making paths and resources

In this section, we address Research Question 1. First, we show the decisions identified for each group as abstracted from their reported solution, the same approach we used in Koretsky et al. [29]. Then we present timelines to show chronologically when each decision was considered and, correspondingly, what material resources supported those decisions. Finally, we present the knowledge resources that the groups activated to support their decision making.

Fig. 3 shows a summary of the major decisions and approaches for each group based on their exam submissions as well as their exam score. The entire group received the same score. Similar to our previous analysis of 117 groups in [29], the groups' decisions varied, with Groups A and C taking more sophisticated engineering approaches and Group D the least sophisticated. Often, the students' perceptions as they completed the exam coincided with a complex, temporally emergent engineering task [85], as Barbara illustrated, "Okay. You know somehow, this seemed a lot more complicated when we started. But I'm glad it's like, less complicated now."

Fig. 4 shows a timeline from the content analysis, in 2.5-minute intervals, illustrating how each group navigated the major decisions (in blue), what tools they shared on the Zoom screen (green), when they completed their virtual experiments (orange), and when they explicitly attended to the required deliverables (orange). Groups A, B, and C considered multiple decisions simultaneously. They also regularly shared materials on the computer screen, most notably Group B who would quickly transition between several resources. Group C spent more time than the other groups considering what form the data would take

Decision performance based on team's submissions. The **bold** decisions are preferred.

Decision	Group A	Group B	Group C	Group D
Mathematical Form	**Linear**	**Linear**	**Linear**	**Linear**
Rate Constant	Interval (k average)	Interval (k average)	**Interval (slope)**	Point
Time	**Lower CI**	k average	**Lower CI**	k average
Reactor Difference	**Paired t-test**	Eyeball	-	-
Exam Score (max 20)	19	13	16	7

Group A:

Approach Reported: The first thing we did was used the simple linear regression equation for the slope of each run done by each person and reactor to find each runs k value. **We then did a paired t test and compared the difference in each person's reactor 1 and 2 and found that the interval included 0 so we are assuming that both reactors come from the same population**. We then took the k values and did a two sided t test with alpha=.05 for the interval. A one sided t test with aplha=.01 to solve for the lowest k and solved for the time with that k value where 70% of the product has been formed.

Group B:

Approach Reported: To determine the values above, we ran 4 experiments on reactor 1 and reactor 2 for a total of 8 experiments. Using the differential equation dCs/dt - -kCs, this equation was solved to get a linear relationship with concentration and time :lnCs = -Kt+c. All of the data for reactor 1 was graphed and a trendline was created and the same procedure was done for reactor 2. The average of these k values was taken, 0.3165. We found the 95% confidence interval by performing a t-test with an alpha of 0.05 and df of 7 using the 8 experiments. **The time was then found using our equation lnCs = -0.3166t+0.5704 where the average inital concentrations were found and multiplied by 30% to show that 70% of the initial concentration was converted. T solved to = 3.80 hours.**

Group C:

Approach Reported: We made an excel with all our values and found the average of them all. We then plotted the ln of those average to make a linear regression and find B1 and B0. We then made a confidence interval for B1(k) and used the upper bound to solve for t. We did this to make sure we used the least amount of time, but still be 95% sure to get 70% of the sugar reacted.

Group D:

Approach Reported: We determined the k values by separating the differential and then integrating both sides and getting the equation: ln(Cs) = -kt + c. We then used the values from the reactors in excel and graphed them with a linear fit. The average values of the slopes found was the k value. We then found the initial concentrations by finding the value of ln(Cs) at time of zero. We multiplied that value by 0.7 and then plugged that new value into the equation found from the graphs to find the t value. We then used the larger t values of the two reactors because that is the time that both reactors would have at least 70%

Fig. 3. Student performance based on their exam submissions. The bold text in the approach reported for groups A and B are discussed in Section 5.3.

and how to approach analysis (~ 20 min) before jumping into experiments. Group D was largely sequential in decision making, took longer to run their experiments, and shared few resources. Group D did refer to a shared GoogleDoc as they worked, which was not shared on the Zoom screen (not shown). However, instead of using it as a shared resource to brainstorm and progress, they appeared to be doing their calculations individually without shared social practices, using the tool to "check answers." In this way, Group D appeared to mirror School World homework practices of each doing their own work rather than the intention of a collaborative, authentic engineering task.

In addition to the material resources identified in Fig. 4, we identified the *knowledge resources* that the groups used. Table 2 provides categorization of these knowledge resources as well as illustrative examples. The most common resource activated was prior class experience, where groups referred both to other courses they had taken as well as activities in the course studied. In addition, the dynamic data provided by the computer tool enabled decision-making based on observation of virtual laboratory data. Finally, students' framing ranged from anticipating instructor expectations, indicating School World, to engineering contextualization, indicating Engineering World. In the examples shown in Table 2, Group B's discourse centers on School World framing, including how the deliverable is "worded," what "they want" (presumably the instructors), and what decisions will maximize points. In contrast, Group A uses Engineering World contextualization, considering how to report results to their bosses and how an operator would respond to a time recommendation with too many significant figures.

5.2. Division of labor and social roles

In this section, we address Research Question 2 by examining the students' social roles within their groups. First, we orient the analysis by providing a quantitative description of each students talk time. We follow that with a qualitative interpretation of the social dynamics based

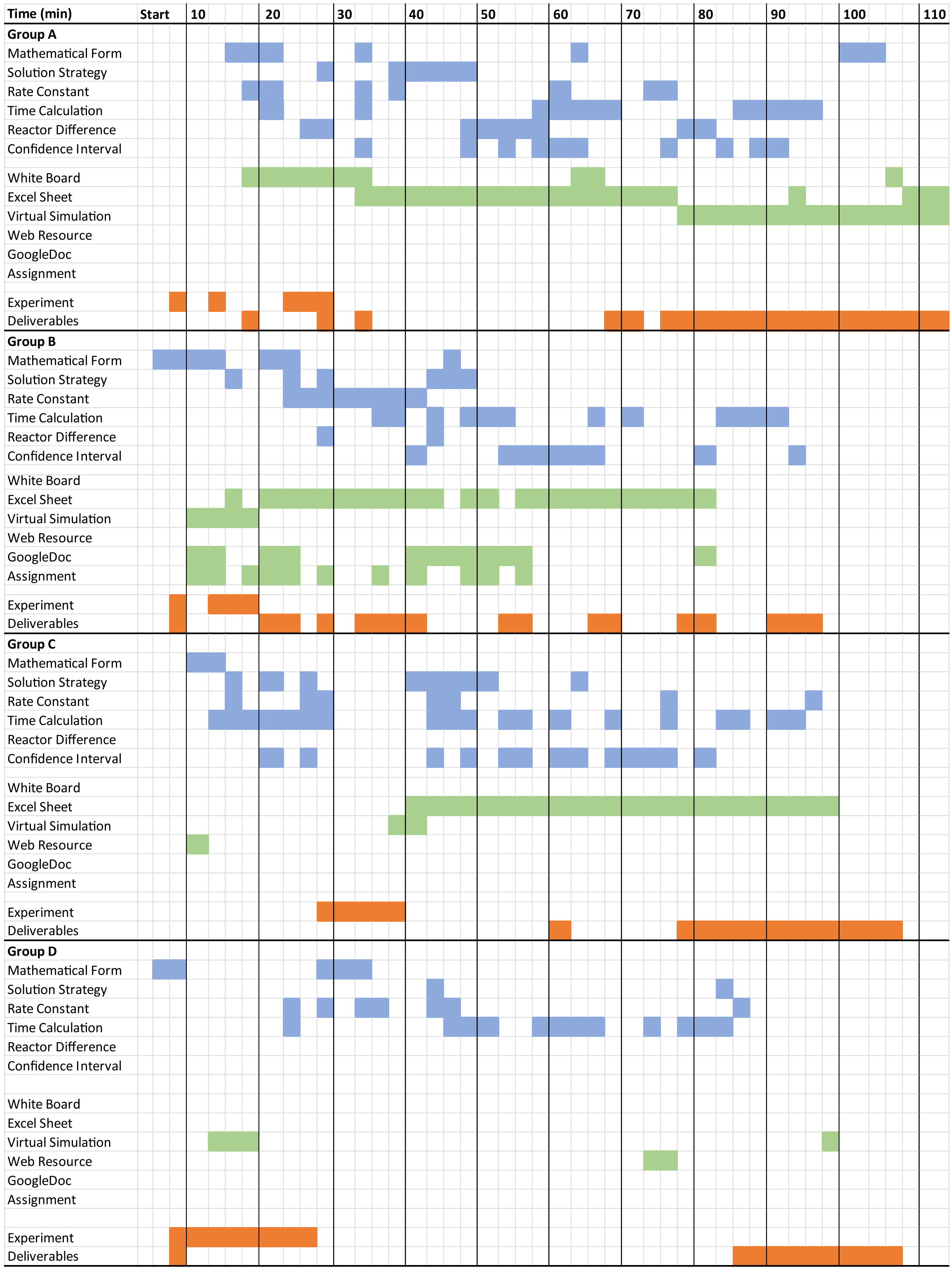

Fig. 4. Timelines of activity generated from the content logging step. Blue bars show the major decisions; green bars show the tools shared on Zoom; orange bars show either the virtual experimentation or attending to deliverables.

on the discourse patterns in the group.

One concern about group exams is social loafing or drafting, where a group member does not engage fully and allows others in the group to carry the workload. To see the degree to which this behavior manifested, we analyzed the division of labor and social roles within the four groups. Fig. 5 shows the distribution of spoken word counts (top) and turns (bottom) for each member of the four groups. To facilitate analysis, the speakers are arranged from the person who spoke the most (Speaker 1) to the one who spoke the least (Speaker 3 in Group D, Speaker 4 in Groups A and B, Speaker 5 in Group C). A turn is defined by uninterrupted speech, no matter how short. On average, each student spoke 2312 words (SD = 1545) while taking 198 turns (SD = 107). As the groups varied in group size, the analysis reported below considers both the number of words spoken and the percentage of the total words that

Table 2
Knowledge resources in the groups' engineering decision making.

Resource	Description	Example
Prior class experience	Students draw on content from both the current course or prior courses to utilize what they have previously learned to generate ideas and make progress	**Group C** within the course: "Um, I was just thinking what if we did like what we did on the homework and that one day in class where we did the line for the lower, uh, confidence interval, and then found on that line where the 70 % was." Connor (01:00:06) **Group D** from other courses: "Do you think maybe there is something with… first order reactions…I feel like we learned something like this in chemistry…" Deetya (01:05:47)
Observation of virtual lab data	Students use the data from their virtual laboratory experimentation in their thinking and decision making	**Group A** Aaron (00:27:10): Yeah, I'm not a fan, but… Oh, okay. Looks like reactor two is consistently slower than reactor one. Angela: Yeah. So reactor- Aaron: Uh oh. Angela: … two doesn't separate. Aaron: How are we going to treat the data separately?
Instructor expectations (School World)	Students look for clues in the assignment or in what the exam is covering to guide their decisions	**Group B** Bailey (01:05:31): Yeah. I don't know. Does it, like, specifically ask for any confidence intervals? Beto: No. Barbara: No. Bailey: Okay. Beto: But it's kinda hinting in the way that it's worded that's the best way. Like, they want it. They want the confidence interval, you know? Bailey: Mm-hmm (affirmative). Beto: 'Cause they're like, how would you report it to them? Like, hmm, give them averages. And then they're like, two points off, that's not the best way.
Engineering contextualization (Engineering World)	Students consider the context from the point of view of an engineer completing a task to guide their decisions	**Group A** Austin (01:27:38): Yeah. The value report to your bosses. Why would you use a different value for your actual calculations and what you report to your bosses? Angela: Well, cause we're reporting two different things. Like, I kind of understand the reasoning, but then we need to make sure that we're all for the reasoning. Austin: Yeah but if- Angela: At least what I'm getting at. Austin: But if, like, another engineer came along down the road, wouldn't they want to actually get within that 70 %? [A little later] Angela (01:35:29): Um, you'd want to round up because adding a little bit more time won't reduce how many of them actually make 70 % [inaudible], barely increasing. Aaron: So, you're just saying- Angela: But for, like, the time we wanna round up if you wanna round. Aaron: Yeah okay. Yeah. We're rounding up of course, right. I'm just saying rounding up for like an added… adding or whatever. And also because you're gonna tell an operator, "can you run that reactor for 6.21 h?" They're gonna punch you in the face!

the group spoke. As became apparent in the qualitative analyses, there were two leadership modes into which the groups assembled – a single-leader mode (Groups A and C) and a dual-leader mode (Groups B and D). In addition to the group leaders, other students took on productive roles consistent with how the group was framing the task, as discussed next. Perhaps coincidentally, the single-leader groups were primarily engaged in Engineering World and the dual-leader groups in School World (e.g., Table 2).

Groups A and C organized their social practices around a clear leader who spoke the most, both of whom were White men: Aaron in Group A (4731 words, 45 % of the words spoken) and Caleb in Group C (3439 words, 53 % of the words spoken). As a fifth-year senior returning from an internship among the other second-year students, Aaron seemed positioned with the authority of seniority, while Caleb appeared to be a confident, outgoing second-year student who took charge. Both Aaron and Caleb immediately took on these leadership roles during the exam, likely a group structure that had been established in the previous five weeks the groups had been working together. Their styles were those of collaborative leaders, guiding their respective groups' progress while simultaneously asking questions that drew others into the problem-solving and engineering decision-making process. For example, while taking the lead on the whiteboard that the group used to collectively work out the mathematical solution, Aaron asked confirming questions to ensure buy in from other group members: "I love drawing things, I can do that. Okay so, dC_S over dt, negative k_{CS}… Are we all with that?" Throughout the exam, Aaron regularly asked clarifying questions, such as "Does this look correct?" and ensured agreement among the group with phrases like "If everyone is good with that," indicating a collaborative, rather than overbearing, leadership style. Similarly, while Caleb drove the work of Group C, such as sharing his work on Excel as a shared screen in Zoom, he called the group in puzzling moments, consistently asking the other members things like "I am just confused about what we should do with this data, if you know what I mean. What do you guys think?" and later, "All right. That seems like - I think we can work with that. Do you guys- what do you guys think? Is that something we agree on?"

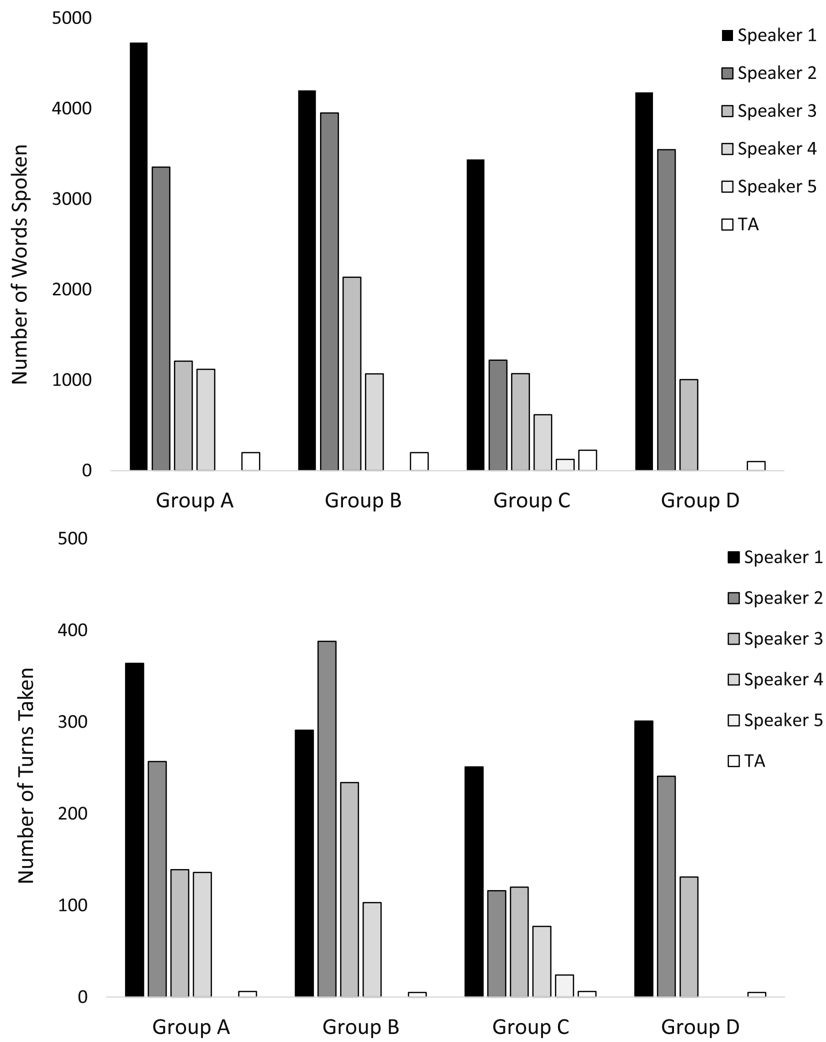

Fig. 5. Words spoken (top) and turn (bottom) counts for each group member. As the group size ranges from three to five, raw totals should not be directly compared.

Groups B and D organized their social practices in a joint (dual) leadership mode. Group B's leadership was shared between Beto, a Latino man, and Baily, a Chinese American woman, and Group D's leadership was shared between Dakota, a White woman, and Deetya, an Indian American woman. As with Aaron and Caleb, if not more so, all four dual leaders encouraged their groups to participate in completing the task. For example, members of Group B had the following exchange about how to organize their work on the shared screen:

Beto (00:22:36): Baily, y-you have it downloaded, right?

Baily: Mm-hmm (affirmative).

Beto: Do you want to do all the work on yours, or do you want us all to…?

Baily: I don't mind doing it. Like, and then we can share, yeah.

Beto: Okay.

Barbara: Yeah. I'm also going to like, follow along on mine, I think.

Baily: Yeah, I was saying like, if that… That usually helps me too, whenever, yeah.

Beto: Might be a good idea, then.

This collaborative approach included calls to help process points of uncertainty and confusion. For example, in Group D, when considering the graphical representation of the sucrose concentration versus time, Deetya shares her thought process with the rest of the group by asking, "What do you guys think about, like, do you think we have to make anything of the y-intercept that we get?"

Often, other group members took on substantive roles to help the group move its work forward, identifying and making decisions. Notably, while not taking the role of leader, several students took on the role of questioner, asking questions that helped the group problematize the task they were working on. For example, to get Group A to consider the Rate Constant decision, Alex asks, "So, as far as like…the final result that we're gonna be presenting … Is it gonna be like, an interval, or is it just gonna be like, just one k-value?" In Group C, Conner problematizes a Solution Strategy decision, "Real quick, wouldn't it be easier if we combined the data for the two reactors rather than doing everything twice?" Similarly, Alex asks about how they should treat the reactor data, "Should we be considering, at all, like the, the individual batches, or like, that each of us ran? And kind of treating that as like, uh, samples?" launching the group into a rich discussion where each student contributes either their own opinion or a follow-up question. Group members also took the lead in more procedural aspects of the task, including demonstrating mastery of Excel (Austin, Bonnie), chemical phenomena (Barbara), and statistics (Angela, Cody). In summary, the activity was often authentic and asset-based in the sense that it allowed individual students to productively contribute to the group's progress in different but complementary ways, demonstrating their individual competencies (e.g., leadership, problematization, technical acuity).

Fourteen students unambiguously contributed to their group's progress while two students did not appear to substantially contribute. Chenxi in Group C spoke only 124 words (24 turns), mostly as she shared her screen when she needed help completing her reactor runs and measurements in the virtual laboratory to contribute to the group's dataset. As a Chinese student in this American classroom, it appeared that a language barrier limited her participation, although her scores on the individual midterms and the ARS questions indicate competency with the material (Table 1). In Group D, Dahlia did not noticeably contribute original ideas. On the contrary, there is evidence she may not have been critically engaging in the task. For example, after approximately 30 min, when the group is considering how to graphically represent sucrose concentration versus time, Dahlia asks the other students, "Wait, you guys are graphing what exactly?" She then proceeds to say, "Okay, because I just graphed it with time and C_S"; however, within the first five minutes, the other students had decided, with apparent consent, that it was best to graph the natural logarithm of the sucrose concentration. This difference exemplifies ways that Dahlia may not have been following along in the problem-solving process of the other students. Dahlia's engagement is puzzling given that she had the highest score of Group D on the individual midterm the preceding day (Table 1).

5.3. Cases of decision-making

In this section, we address Research Question 3, how the students' framing and group dynamics influence their collaborative decision making. As Fig. 4 shows, the groups' decision-making processes were interwoven and complex. To illustrate this, we select examples of how Groups A and B made the engineering decisions highlighted in bold in Fig. 3. The groups engaged in different figured worlds (Table 2). While Group A's decision was productive and Group B's less so, we argue that both cases illustrate how an individual's ideas and thinking become entangled within the group practice. From these cases, we propose that authentic assessment may need to consider the unit of the group in a way that extends beyond the isolated contributions of each individual member.

5.3.1. Reactor differences decision (Group A)

We characterize the decision made by Group A to use the statistical methods covered in the course to analyze the difference between the two reactors. The only other group to consider reactor difference was Group B, which they merely eyeballed and chose to ignore. In fact, Group B's data was drawn from different values of k with k_1=0.28 and k_2=0.32, so a more systematic analysis would have been appropriate.

Group A largely took on the role of engineers in Engineering World, as illustrated by Austin and Aaron's reasoning in Table 2 (Engineering contextualization). However, Angela struggled with the ill-defined problem, leading her to make several bids for School World. She appears most concerned that the answers and explanations the group is providing would be satisfactory for the instructional team who created the exam (Instructor expectations). When the group considers the decision on *Solution Strategy,* Angela is quick to suggest emailing the instructional team for guidance on how to approach data from each group member's runs:

> Angela (40:35): That's what I'm wondering. Should we email them [the instructional team] and like, ask like, what the assumptions are for our reactors? So like, are they the same reactor that each of us are running, at a different time? Or are they actually different reactors?

Despite others in the group being resistant to this suggestion (Aaron saying, "Uh, they probably won't say anything," and Austin adding, "They'll probably say something along the lines of, 'Ha, ha, ha. That might be a good assumption to make. Ha, ha.'"), Angela emails them anyway. She receives the response of "It is up to your group to decide how to calculate the parameters asked of you. There is more than one way to do so correctly." She interprets this response in a way that aligns with a School World approach; as long as they clearly write down every assumption they make, they should be justified in their answer.

> Angela (50:01): So, I think as long as we write down every assumption we make, um … So, if we want to assume both reactors are from the same population … um … we can then put Reactor 1 and 2 together. Um …

Rather than prioritizing the choice of the most productive engineering approach to the problem, she prioritizes stating the justification for any approach they do pick, placing more value on the words and requests of the instructors rather than holding herself accountable for engineering decisions. She appears to think the "game" is that we can choose any assumption that we want as long as we justify it; in other words, she does not see any consequences to her assumptions.

However, Aaron gently leads the group away from the arbitrary assumption ("Um … Let's get like, the average"), which leads them to calculate confidence intervals (CI) for both reactors to compare them. As they make progress on the calculations, Angela jumps back in, suggesting they do a paired *t*-test and see if the CI crosses zero.

> Alex (00:54:54): Yes [25 s pause]. Wait, so what we did here justifies that both reactors can be treated as just one gigantic, like, data set? Can we make that justification and then go back to the linear regression and just kind of use that for, just additional, um … like, support … for whatever value or interval we get?
> Aaron: Um … [crosstalk]
> Angela: You do the difference of them? 'Cause you do the difference of them, and the difference in the confidence interval includes zero. And then we can say that they … At least we're 95 % confident-
> Alex: Mm-hmm (affirmative).
> Angela: … Or at least cannot throw out, uh, that they could be from the same population.
> Alex: Mm-hmm (affirmative).
> Aaron: Okay, the difference of, uh …
> Angela: So each of the runs, um, take your Reactor 1 run minus your Reactor 2 run. There'll be a difference-
> Angela: Oh, okay.
> Angela: … And take that difference, and then do the confidence interval on that. Um … and then if zero's in … If zero's not included, we can assume … We can be 99, 95 % confident with the current t-value. Uh-
> Alex: Ah-
> Angela: … that they are from different populations. Or that there is enough of a difference that …
> Austin: So we should do that with the second reactor, too. Like, all eight data sets? Is that what we're doing?
> Angela: Yeah, so you're taking each person's Reactor 1 minus your Reactor 2. Or, you can do it the other way around. Um … then there'll be a difference, um …
> Aaron: Okay, and then we just … Then we can just do this … Is that lined up here? It includes zero, so …
> Angela: Okay, so are we all conf- at least willing to say that we are not confident enough that they are different?
> Aaron: We have been unable to show that they're different, that's for sure. That's for damn sure!

In proposing this idea, Angela demonstrates deep statistical understanding with appropriate language ("Okay, so are we all conf- at least willing to say that we are not confident enough that they are different?") which leads the group back to her original suggested assumption, but now in a systematic way more appropriate for engineering practice that is supported by data and statistical analysis. Finally, once resolved, the group shows quite positive affect, all smiling and taking a moment to joke back and forth.

We ask, could Angela's contribution have occurred without the group's framing? We argue there is a subtle but important difference between being able to exhibit competencies *by yourself* and being ready to contribute within the context of group practice. Among other purposes, assessments are a way to show students the valued practices of a community. In this way, Group A's practice positioned Angela to make a substantial knowledge-based contribution. However, students are not always positioned that way within group practice as illustrated next.

5.3.2. Production time decision (Group B)

Group B largely took on the role of students completing an exam assignment in School World. In Table 2 (instructor expectations), Beto does not foreground what makes the most sense from a situated engineering perspective, but instead emphasizes the instructor's expectations relative to the exam scoring. Indeed, the group shows strong inclinations to School World decision-making throughout the task. They check on the task's deliverables, more than the other groups (Fig. 4), from the perspective of not missing points on an exam.

However, they demonstrate engineering reasoning in one case as they work to decide what reactor time to recommend to production. After having obtained a time (3.8 h) based on a point estimate of the reaction constant *k*, Beto proposes considering an interval estimate.

Beto (01:24:47): You know what … Should we change our time … to the highest value of k? Instead of the average? Because it needs a … Like they can't have less than 70 % converted, but you can have a little bit more. So, if we use one of the extremes of k, and we get more time, then that would probably be the best time because at 3.8 [hours] we should be averaging, like half of the reactions, or less. If we're assuming that none of them are actually on the line, if that makes sense.
Barbara: Oh, I see what you mean.
Bailey: Mm-hmm (affirmative).
Beto: So…
Barbara: I mean … I don't know.
Bailey: Well, will that change if you have … 'Cause like when we solve for it, we put in the 30 %, so wouldn't that guarantee that it, it's, that it converts to 70 at least? At least with-
Beto: No.
Bailey: … our k value.
Beto: No, because, like we're doing everything in terms of averages. So, we're assuming that realistically it's probably never gonna be like exactly on the [production] line that we're expecting. We're always gonna have at that point, at 3.8 h, ah, we're always gonna have points above and points below our-

Beto considers the engineering context, "on the line," to argue that they should account for reactor variation in deciding about production time, a major point of the assessment. Next, Beto continues with this engineering contextualization, stating, "Then shouldn't we do that? 'Cause we're the ones that they're consulting about it, and they're already telling us that they want at least 70 %." Rather than framing the decision in terms of instructor expectations as they have done up to this point, here Beto positions the group as engineering consultants. Bonnie agrees, and Barbara agrees reluctantly as well. After some discussion of whether they should use the higher or lower limit of the confidence interval, however, Barbara renews her objection, stating a concern for the time remaining in the exam. After Bonnie suggests support for this concern, Beto relents, and the group reports the average *k* rather than the more appropriate lower confidence bound.

Barbara (01:28:04): Yeah, I think it would be better. But, I think we still, like then that would change everything that we did. And I think-
Beto: I mean …
Barbara: … um, it's kind of, it'll be okay. I think.
Beto: Does it change everything? I mean it would just change the confidence. No, it doesn't even change the confidence interval, it just changes my last-minute calculations on the Excel sheet.
Barbara: It would change the model and the time, yeah.
Beto: Yeah.
Bonnie: Yeah, it would've changed, it would change questions 2 and 4.
Beto: Okay. I'm cool with it.

Later, they revisit this decision, realizing they had 20 min remaining to complete the task:

Barbara (01:37:48): We actually might finish this earlier than we've finished most of them [class studio activities] (laughs).
Bonnie: (Laughs).
Beto: I don't know, I mean, this is pretty normal 1:40.
Barbara: Yeah, I guess so, but it's, you know…
Beto: Yeah,
Barbara: Maybe a couple minutes early.
Beto: Yeah, I'm glad you talked me out of it (laughs).
Bailey: (Laughs).
Barbara: Yeah, I don't think we would've had enough time.
Beto: Yeah. I mentioned it, I said if we had more time, we likely would've used-
Barbara: Mm-hmm (affirmative).
Beto: … the smaller one.

While Beto acknowledges the preferred solution ("the smaller one.") and, realistically, the group had plenty of time to complete the calculation, the group's practice led to a prioritization of completion despite their engineering judgment.

These two cases illustrate the juxtaposing ways that the individual's thinking becomes entangled in the group practice. Angela's inclination was to appeal to the instructor's expectations and justify assumptions superficially without attending to engineering substance. Only when others steered the group to account for the engineering context was Angela able to demonstrate her deep understanding of the statistical analysis. Likely, on her own, she would have stumbled on the task's open-endedness. In contrast, Beto sought to account for process variation in recommending a reactor time. However, this idea emerged within a group context that was oriented towards the assessment metrics within the fixed-time setting of an exam. Ultimately, that orientation leads others in the group to rebuff Beto's idea in favor of the seemingly more expedient choice. These cases illustrate that while individual competencies can be characterized in this group exam, the group outcome can lead to more (or less) than the sum of the individuals. We propose that this aspect, as captured by the exam, is *group practice* and is worthy of an assessment outcome in addition to each individual's knowledge and skills.

6. Discussion

This study used interaction analysis of video data of four groups to explicate how a computer-based exam served as an assessment tool in a core undergraduate engineering course in the US. By including process variation, measurement variation, multiple possible solution paths, and the need to use engineering judgment, the reactor simulation task provided students with an industrially situated activity that aligned with conceptions of authentic learning and authentic assessment [26,27]. The detailed qualitative analysis of four focal groups allowed insight into processes that were not visible in the prior larger-scale analysis of 117 groups [29].

For some groups the assessment was authentic in the sense that they framed the task in Engineering World, employing knowledge practices that reflected the way engineering is conducted in real life [86]; however, other groups primarily used School World reasoning attending primarily to instructor expectations [65]. Additionally, the task provided students with the opportunity to assume different social roles [40, 37] as multiple perspectives became a valuable resource in the groups' decision-making process [38]. In this way, the assessment pushed students beyond knowledge reproduction and lower-order thinking processes typical of many engineering school assessments [87,27]. Such assessment practices are becoming imperative in the 21st century, where AI tools can readily perform lower-level tasks.

Assessment plays two key roles in the university curriculum. First, it measures students' individual mastery of course content. This aspect provides feedback on the degree to which the instruction meets the learning objectives and ensures that students possess the proper credentials that their degree represents [1]. From a credentialing perspective, group exams have been criticized since some students may draft and not fully engage in the task [57,59]. Of the sixteen students observed here, most engaged deeply with the task, albeit in different ways. However, two students did not, one apparently due to a language barrier, while the other appeared to have trouble following the group's solution process. This lack of participation could be an issue for credentialing. One resolution is to include a mix of individual and group assessments, as was done in the course studied here. Three groups had members who were impressively proficient at the procedural aspects of the work once the group had made its engineering decision. This aspect is essential in completing engineering tasks. Such proficiency may have been fortified through their accountability in the individual exams and/or through other aspects of the course [71]. Group practice should not be developed to the detriment of such technical proficiency but rather in conjunction with it.

Second, although less often acknowledged, assessments shape how students understand what counts as knowledge and how knowledge and social practices are utilized in the profession [9]. In other words, the kind of knowledge students gain and how they learn to apply it are influenced by the assessments they complete. To this point, a strength of this computer-based exam was the clear cases of group practice [38,48], where the work of the individuals became entangled in the group's approach. Group A was largely in Engineering World and was able to support Angela in ways she likely would not have realized on her own. In contrast, Group B was largely in School World and rebuffed Beto's bid to account for variation in their recommended production time. If we view assessments as a tool to guide students toward the professional practices expected in engineering, then the group outcome and the nature of their group practice become paramount. That is, if we want engineering students to develop proficiency in decision-making, we need to attend to the roles in which they are socially positioned as well as how the group takes up the task. Linked to such group practices are critical 21st-century skills, including communication, collaboration, critical thinking, problem-solving, self-awareness, and self-confidence [26]. Here, we do not discount individual assessment as an important slice of the assessment pie but advocate that group assessment is needed too. With the group as a unit of analysis, more complex, interactive group dynamics come into play, aspects that we do not have access to when we assess individuals.

While we were able to unpack these aspects in this qualitative study, it required laborious effort even for only a small proportion of the groups who took the exam. An outstanding assessment question is how do we assess group practice when it is more than the sum of the individuals? This becomes imperative because group practice is integral to how engineering work is typically done [35] and how professionals are evaluated [37]. As educators, we want to give the message that such group practice is valued and cultivate productive group practice. One possibility is to pursue tools that embed generative AI [88,89] into the type of computer-based authentic assessment described here, providing information about the groups' interactions with the tool [6] leading to an evaluation of the group practice. It would be straightforward to such a tool generated word and turn counts (Fig. 5) or the degree that material tools were shared (Fig. 4). More challenging would be to identify the complementary roles exhibited, for example, by Group A. Research is needed to further characterize group practice and develop corresponding assessment strategies.

This qualitative study has a small sample size, making it inappropriate to draw conclusions about causality from the results. While limited to the detailed analyses of only four focal groups, we observed some patterns in how groups assumed social roles and made sense of the task that were interesting and may be fruitful for future investigations. In the cases studied here, leadership modes were associated with group demographics. Both groups assuming the single-leader organization had White men who took up that role. In contrast, the two groups comprised entirely of underrepresented students took on a dual-leadership organization. In addition, the former groups tended to take up the roles of engineers working on a task in Engineering World, while the latter were oriented as students completing an assignment in School World. These patterns are consistent with other studies we have conducted [38,55,65,90]. Taken together, we might conjecture that the minoritized students are more attuned to hierarchical social patterns and bring with them a social asset that tends towards a distributed leadership approach. At the same time, they could more regularly face challenges to their legitimacy as engineering students, making them less likely to reframe higher-stakes school assignments as a hypothetical engineer in practice. These relationships between social identity and engineering learning would be a fruitful avenue for future research.

There are several limitations to this study. This study was conducted during Spring quarter 2020 at the height of emergency remote teaching due to the onset of the COVID-19 pandemic. Both the instructors and students were scrambling to adjust to remote teaching as well as greater disruptions in their lives. These times were, undoubtedly, unusual, and the interactions reported here should be considered with this context in mind. However, the groups' responses and decision-making choices during this term were consistent with the earlier data reported of 117 groups before the pandemic [29]. The core focus of the authentic assessment to elicit decision making strategies still appears valid. In fact, in the rapidly shifting landscape of STEM instruction, such as the release of ChatGPT in 2022, we argue that the pandemic provides a natural stress test for the type of authentic assessment reported here and in Koretsky et al. [29]. While it certainly limited groups' interactions, the shift to Zoom recordings provided a practical advantage as well; it facilitated audio quality relative to video data collection in a noisy in-person classroom. Finally, Groups A and D were analyzed collaboratively while Groups B and C were analyzed independently by the first author due to the second author's limited availability. These latter two groups were essential to verify our claims about single (Groups A and C) and dual leadership (Groups B and D) modes. We had developed a stable codebook by the time of the transition, but there could be discrepancies, and the results should be considered with this limitation in mind.

As a small-sample size, qualitative study, we do not claim generalizability, but rather we aim for transferability, where others can build on the findings from the computer-based, authentic group assessment described here. This authentic exam provided a problem where students had the opportunity to toggle between fundamental principles, as manifest in the form of reaction kinetics, and statistics, as they confronted a system with process and measurement variation. In principle, this type of group exam is not limited to the context described in this study; rather, this form of computer-based authentic assessment could be delivered in a variety of engineering and science courses where the instructor wants to expand the scope of the assessment to include decision making and group practice. However, the creation of the task requires disciplinary expertise, and the development of the associated technology is labor intensive. More research is needed to identify ways that this style of authentic assessment can be more accessible for instructors who do not have those resources. Alternatively, educational technology developers could focus on developing tools such as the sugar reactor virtual laboratory described here to enable group exams for other contexts. Regardless of how these exams are developed, we believe a mix of individual and group exams should be used to address the limitations of individual accountability as demonstrated by two of the fourteen students in this study. However, developing and identifying competencies of group practice may be more complicated, as both students who showed minimal engagement in the group exam performed well on the individual exam the day before.

While generative AI was not firmly grounded in the classroom ecology when this study was conducted, with the public release of

ChatGPT in late 2022 and the rapid emergence of other generative AI tools, university STEM learning environments have entered a period of substantial disruption. This shift has required instructors and institutions to reconsider assessment, academic integrity, feedback, student support, and the role of human sensemaking in learning [91]. These tools are increasingly able to answer even the most challenging single correct answer assessment questions that have historically provided the means to test competency in STEM. Correspondingly, it is not surprising that students would find difficulty seeing the value in answering this type of questions on their own and, by extrapolation, in courses that focus solely on this competency. While preliminary, this study suggests ways to shift the assessment landscape in this disruptive context. Using a computer-based technology, we have foregrounded open-ended decision making over procedural fluency or conceptual understanding. We argue this approach exemplifies a paradigmatic shift that is paramount to keep assessments relevant in the age of generative AI.

7. Conclusion

In an era where generative artificial intelligence is reshaping professional practice, engineering assessments must evaluate not only individual knowledge but also authentic group work. This study investigates how second-year engineering students in the United States engaged in a computer-based group exam situated in a Unity 3D simulation of a candy factory. Four teams were video recorded as they collected reactor data and made engineering decisions during a two-hour session. Interaction analysis of videos and transcripts, supported by exam responses and course scores, revealed that higher-performing groups shared digital tools and shifted fluidly between decision points. In contrast, the lowest-performing group rarely shared tools and followed a rigid, sequential process. Leadership patterns varied, but most students actively contributed; however, two students participated minimally, raising questions about using such exams for credentialing purposes. Framed through the lens of figured worlds, findings show that individual actions become entangled in group dynamics, suggesting the need to treat the group, not just the individual, as a core unit of analysis in engineering assessment.

CRediT authorship contribution statement

Milo D. Koretsky: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Irene Pantekidis:** Writing – review & editing, Writing – original draft, Formal analysis.

Declaration of competing interest

None of the authors have a conflict of interest with this study.

Acknowledgements

The work reported here was supported in part by grants from the National Science Foundation (EEC 1519467) and Tufts University through the Faculty Research Awards Committee. The views represented here are those of the authors. The authors thank Jeffrey Nason and Samuel Gavitte for their feedback.

Appendix A. Assessment rubric

Table A1

Table A1
Rubric for group exam assessing social practices and decision-making.

Problem	Score	Low = 0	Medium	High = max	Possible
Group collaboration and		**0 pts** The work and approach are dominated by one individual	**2 pts** The work is mostly distributed with all having a chance to participate.	**4 pts** All team members are actively participating, listening to others and sharing their ideas	4
Engagement		**0 pts** Substantial periods of disengagement	**2 pts** Not all team members are active; or slight periods of disengagement	**4 pts** The team is actively collaborating and working towards a solution the entire time.	4
k		**0 pts** Single value reported	**2 pt** CI reported but does not reasonably match range it should be	**4 pts** CI reported **and** reasonably matches range it should be	4
Production time		**0 pts** Average k used for time	**2 pts** CI low used for time	**3 pts** … CI low and reported and time reasonably fits spreadsheet	4
Justification		**0 pts** No justification provided	**1 pt** Significant elements missing	**2 pts** Solution process well described and justified	2
Reactor Difference		**0 pts** Reactor difference not considered	**1 pt** Considered but no statistical analysis	**2 pt** Difference confirmed or negated with appropriate statistical analysis	2
Total					**20**

Appendix B. Codebook

Table B2

Table B1
Atlas.ti codes.

Code Category	Code	Description
Social	Collaborative work	Students working together to come up with solution
	Confirming question	Student asks follow-up question to ensure understanding of concept or proposed problem-solving method
	Negotiation	Students negotiate amongst each other and either land on a shared decision or agree to disagree
	New idea introduced	Student introduces an idea that has not been previously considered or that is different from the present group thinking
	Proposes analysis method	Student proposes a strategy for analyzing the present part of the problem
	Regulatory	Student brings up a different idea or a concern with current group thinking
	Compromise	Students come to a compromise on an answer or strategy after group discussion
Figured Worlds	School world	Student's thinking or ideas align with figured school world
	Engineering world	Student's thinking or ideas align with figured engineering world
	Bid for figured world rejected	Student rejects school world or engineering world thinking or ideas of other group members
Decision	Confidence interval	To determine the rate constant, k, the group provides an interval based on calculated confidence interval
	Linear regression	Linear model used to show the dependence of the concentration on time
	One tailed *t*-test	One tailed *t*-test used to determine statistical significance of difference between two reactors
	Relate k value to time	Time was calculated based on the reported k value
	Significant figures	Decision surrounding number of significant figures to use in reported answers
	Solving equation	Solving equation to determine value needed for report or for subsequent calculations
	Treat reactors the same/treat reactors differently	Treating all reactors the same or differently based on statistical analysis
	Treat all samples as one/treat all samples differently	Treating all samples the same or differently for subsequent calculations
Episode	Episode	Change of episodes with the introduction of a new idea, the use of a new strategy, or the transition from one figured world to another.

Appendix C: Sample coded transcript

Fig. C1 shows a screenshot of an *Episode* from Group A coded in Atlas.ti. In this episode at 27:10, Speaker 1 (Aaron), observing the data they collected, notes "reactor two is consistently slower than reactor one" instigating an interaction with Speaker 4 (Angela). This move is coded as *Social: New Idea Introduced* (See Appendix B). From this idea, Speaker 1, asks "How are we going to treat the data separately," coded as *Decision: Treating Reactors Differently*."Speaker 4 responds to this move by suggesting they look at the assignment instructions, which both Speaker 1 and Speaker 2 (Alex) take up. This initial approach looks for clues of the answer from the instructions (instructor) rather than using their engineering knowledge and is therefore coded as *School World*.

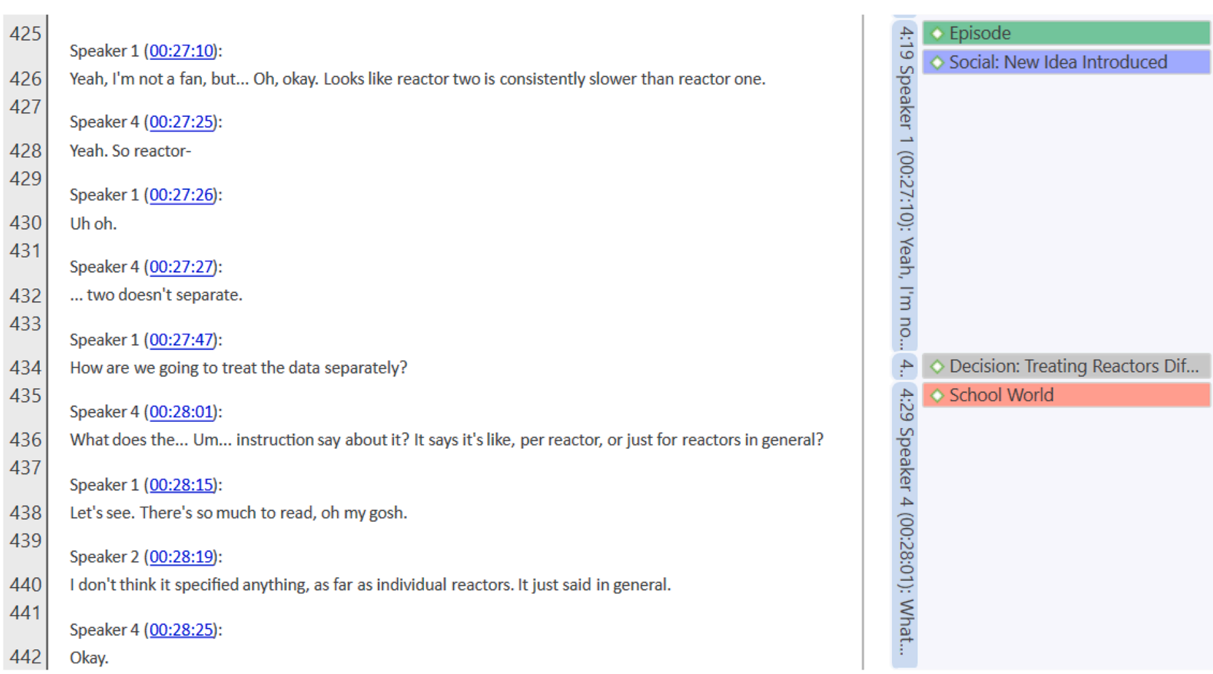

Fig. C1. Coded transcript excerpt from group A.

References

[1] Pellegrino JW, Wilson MR, Koenig JA, Beatty AS, National Academies NRC. Developing Assessments for the Next Generation Science Standards. Washington, D.C: National Academies Press; 2014. http://www.nap.edu/catalog.php?record_id=18409.
[2] Shavelson RJ, Baxter GP, Pine J. Performance assessment in science. Appl Meas Educ 1991;4(4):347–62. https://doi.org/10.1207/s15324818ame0404_7.
[3] Shute VJ, Leighton JP, Jang EE, Chu M-W. Advances in the science of assessment. Educ Assess 2016;21(1):34–59.
[4] Asher MW, Hartman JD, Blaser M, Eichler JF, Carvalho PF. The promise of mastery-based testing for promoting student engagement, self-regulated learning, and performance in gateway STEM courses. Comput Educ 2025;237:105387. https://doi.org/10.1016/j.compedu.2025.105387.
[5] Daradoumis T, Puig JMM, Arguedas M, Liñan LC. Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming. Comput Educ 2019;128:159–70. https://doi.org/10.1016/j.compedu.2018.09.021.
[6] He S, Cui Y. A systematic review of the use of log-based process data in computer-based assessments. Comput Educ 2025;228:105245. https://doi.org/10.1016/j.compedu.2025.105245.
[7] Schiller R, Fleckenstein J, Mertens U, Horbach A, Meyer J. Understanding the effectiveness of automated feedback: using process data to uncover the role of behavioral engagement. Comput Educ 2024;223:105163. https://doi.org/10.1016/j.compedu.2024.105163.
[8] Biggs J. Constructive alignment in university teaching. HERDSA Rev High Educ 2014;1:5–22. In: editor.
[9] Boud D. Assessment and the promotion of academic values. Stud High Educ 1990;15(1):101–11. https://doi.org/10.1080/03075079012331377621.
[10] Jordan ME, Babrow AS. Communication in creative collaborations: the challenges of uncertainty and desire related to task, identity, and relational goals. Commun Educ 2013;62(2):210–32. https://doi.org/10.1080/03634523.2013.769612.
[11] Gratchev I, Howell S, Stegen S. Academics' perception of final examinations in engineering education. Australas J Eng Educ 2024;29(1):20–9. https://doi.org/10.1080/22054952.2023.2284484.
[12] Lujan HL, DiCarlo SE. Too much teaching, not enough learning: what is the solution? Adv Physiol Educ 2006;30(1):17–22. https://doi.org/10.1152/advan.00061.2005.
[13] Price AM, Kim CJ, Burkholder EW, Fritz AV, Wieman CE. A detailed characterization of the expert problem-solving process in science and engineering: guidance for teaching and assessment. CBE—Life Sci Educ 2021;20(3):ar43. https://doi.org/10.1187/cbe.20-12-0276.
[14] Gernsbacher MA, Soicher RN, Becker-Blease KA. Four empirically based reasons not to administer time-limited tests. Trans Issues Psychol Sci 2020;6(2):175. https://doi.org/10.1037/tps0000232.
[15] Stanger-Hall KF. Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE—Life Sci Educ 2012;11(3):294–306. https://doi.org/10.1187/cbe.11-11-0100.
[16] De-La-Fuente-Valentín L, Pardo A, Kloos CD. Addressing drop-out and sustained effort issues with large practical groups using an automated delivery and assessment system. Comput Educ 2013;61:33–42. https://doi.org/10.1016/j.compedu.2012.09.004.
[17] Borges B, Foroutan N, Bayazit D, Sotnikova A, Montariol S, Nazaretsky T, Banaei M, Sakhaeirad A, Servant P, Neshaei SP. Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. Proc Natl Acad Sci USA 2024;121(49):e2414955121. https://doi.org/10.1073/pnas.2414955121.
[18] Frenkel ME, Emara H. ChatGPT-3.5 and-4.0 and mechanical engineering: examining performance on the FE mechanical engineering and undergraduate exams. Comput Appl Eng Educ 2024;32(6):e22781. https://doi.org/10.1002/cae.22781.
[19] Wang KD, Burkholder E, Wieman C, Salehi S, Haber N. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. Front Educ (Lausanne) 2024;8:1330486.
[20] Gruenhagen JH, Sinclair PM, Carroll J-A, Baker PR, Wilson A, Demant D. The rapid rise of generative AI and its implications for academic integrity: students' perceptions and use of chatbots for assistance with assessments. Comput Educ: Artif Intell 2024;7:100273. https://doi.org/10.1016/j.caeai.2024.100273.
[21] Huang D, Hash N, Cummings JJ, Prena K. Academic cheating with generative AI: exploring a moral extension of the theory of planned behavior. Comput Educ: Artif Intell 2025;8:100424. https://doi.org/10.1016/j.caeai.2025.100424.
[22] Lyu W, Zhang S, Chung T, Sun Y, Zhang Y. Understanding the practices, perceptions, and (dis) trust of generative AI among instructors: a mixed-methods study in the US higher education. Comput Educ: Artif Intell 2025;8:100383. https://doi.org/10.1016/j.caeai.2025.100383.
[23] Ortiz-Bonnin S, Blahopoulou J. Chat or cheat? Academic dishonesty, risk perceptions, and ChatGPT usage in higher education students. Soc Psychol Educ 2025;28(1):113. https://doi.org/10.1007/s11218-025-10080-2.
[24] Susnjak T, McIntosh TR. ChatGPT: the end of online exam integrity? Educ Sci (Basel) 2024;14(6):656. https://doi.org/10.3390/educsci14060656.
[25] McArthur J. Rethinking authentic assessment: work, well-being, and society. High Educ (Dordr) 2023;85(1):85–101. https://doi.org/10.1007/s10734-022-00822-y.
[26] Sokhanvar Z, Salehi K, Sokhanvar F. Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: a systematic literature review. Stud Educ Eval 2021;70:101030. https://doi.org/10.1016/j.stueduc.2021.101030.
[27] Villarroel V, Boud D, Bloxham S, Bruna D, Bruna C. Using principles of authentic assessment to redesign written examinations and tests. Innov Educ Teach Int 2020;57(1):38–49. https://doi.org/10.1080/14703297.2018.1564882.
[28] Wiggins G. The case for authentic assessment. Pract Assess Res Eval 1990;2(1):2. https://doi.org/10.7275/ffb1-mm19.
[29] Koretsky MD, McColley CJ, Gugel JL, Ekstedt TW. Aligning classroom assessment with engineering practice: a design-based research study of a two-stage exam with authentic assessment. J Engin Edu 2022;111(1):185–213. https://doi.org/10.1002/jee.20436.
[30] Chang C-W, Lee J-H, Wang C-Y, Chen G-D. Improving the authentic learning experience by integrating robots into the mixed-reality environment. Comput Educ 2010;55(4):1572–8. https://doi.org/10.1016/j.compedu.2010.06.023.
[31] Hu A, Liu Q, Daniel B. Digital technologies in authentic assessment in higher education: a systematic literature review and narrative synthesis. Sage Open 2025;15(3):21582440251357198. https://doi.org/10.1177/21582440251357198.
[32] Nieminen JH, Bearman M, Ajjawi R. Designing the digital in authentic assessment: is it fit for purpose? Assess Eval High Educ 2023;48(4):529–43. https://doi.org/10.1080/02602938.2022.2089627.
[33] Jonassen D, Strobel J, Lee CB. Everyday problem solving in engineering: lessons for engineering educators. J Eng Educ 2006;95(2):139–51. https://doi.org/10.1002/j.2168-9830.2006.tb00885.x.
[34] Jonassen DH. Instructional design models for well-structured and III-structured problem-solving learning outcomes. Educ Technol Res Dev 1997;45(1):65–94. https://doi.org/10.1007/bf02299613.
[35] Vincenti WG. What Engineers Know and How they Know It: Analytical Studies From Aeronautical History. Baltimore: MD: Johns Hopkins University Press; 1990.
[36] Murray JK, Studer JA, Daly SR, McKilligan S, Seifert CM. Design by taking perspectives: how engineers explore problems. J Eng Educ 2019;108(2):248–75. https://doi.org/10.1002/jee.20263.
[37] Trevelyan J. The making of an expert engineer. Boca Raton: FL: CRC Press; 2014.
[38] Chindanon K, Koretsky MD. Group practice in engineering: productive interactions during a realistic, open-ended task. Stud Engin Edu 2023;4(2):144–75. https://doi.org/10.21061/see.112.
[39] Trevelyan J. Transitioning to engineering practice. Eur J Eng Educ 2019;44(6):821–37. https://doi.org/10.1080/03043797.2019.1681631.
[40] Bucciarelli LL. An ethnographic perspective on engineering design. Des Stud 1988;9(3):159–68. https://doi.org/10.1016/0142-694x(88)90045-2.
[41] Lotan RA. Group-worthy tasks. Educ Leadersh 2003;60(6):72–5.
[42] Johnson DG. Rethinking the social responsibilities of engineers as a form of accountability. In: Michelfelder D, Newberry B, Zhu Q, editors. Philosophy and engineering: exploring boundaries, expanding connections. 26. Edited by Cham: Springer; 2016, 20168598.
[43] McGowan VC, Bell P. Engineering education as the development of critical sociotechnical literacy. Sci Educ (Dordr) 2020;29(4):981–1005. https://doi.org/10.1007/s11191-020-00151-5.
[44] Pawley AL. What counts as engineering: towards a redefinition. In: Baillie C, Pawley A, Riley D, editors. Engineering and social justice: In the university and beyond. West Lafayette, IN: Purdue University Press; 2008. p. 59–85. https://doi.org/10.2307/j.ctt6wq5pf.7.
[45] Esmonde I, Langer-Osuna JM. Power in numbers: student participation in mathematical discussions in heterogeneous spaces [Article]. J Res Math Educ 2013;44(1):288–315. https://doi.org/10.5951/jresematheduc.44.1.0288.
[46] Henderson TS. Understanding access to learning opportunities in collaborative projects: gendered social hierarchies in student teams. Stud Eng Educ 2023;4(1):90–114. https://doi.org/10.21061/see.101.
[47] Wieselmann JR, Dare EA, Ring-Whalen EA, Roehrig GH. I just do what the boys tell me": exploring small group student interactions in an integrated STEM unit. J Res Sci Teach 2020;57(1):112–44. https://doi.org/10.1002/tea.21587.
[48] Medina R, Stahl G. Analysis of group practices. In: Cress U, Rosé C, Wise A, Oshima J, editors. International handbook of computer-supported collaborative learning. Cham: Springer; 2021. p. 199–218.
[49] Menekse M, Purzer S, Heo D. An investigation of verbal episodes that relate to individual and team performance in engineering student teams. Int J STEM Educ 2019;6(1):7. https://doi.org/10.1186/s40594-019-0160-9.
[50] Oner D. Tracing the change in discourse in a collaborative dynamic geometry environment: from visual to more mathematical. Int J Comput Support Collab Learn 2016;11(1):59–88. https://doi.org/10.1007/s11412-016-9227-5.
[51] Lämsä J, Hämäläinen R, Koskinen P, Viiri J. Visualising the temporal aspects of collaborative inquiry-based learning processes in technology-enhanced physics learning. Int J Sci Educ 2018;40(14):1697–717. https://doi.org/10.1080/09500693.2018.1506594.
[52] Barron B. When smart groups fail. J Learn Sci 2003;12(3):307–59. https://doi.org/10.1207/s15327809jls1203_1.
[53] Jordan ME, McDaniel Jr RR. Managing uncertainty during collaborative problem solving in elementary school teams: the role of peer influence in robotics engineering activity. J Learn Sci 2014;23(4):490–536. https://doi.org/10.1080/10508406.2014.896254.
[54] Kittleson JM, Southerland SA. The role of discourse in group knowledge construction: a case study of engineering students. J Res Sci Teach 2004;41(3):267–93. https://doi.org/10.1002/tea.20003.
[55] Hirshfield L, Koretsky MD. Gender and participation in an engineering problem-based learning environment. Interdiscip J Probl-Based Learn 2018;12(1):2. https://doi.org/10.7771/1541-5015.1651.
[56] Langer-Osuna JM, Munson J, Gargroetzi E, Williams I, Chavez R. So what are we working on?": how student authority relations shift during collaborative

mathematics activity. Educ Stud Math 2020;104(3):333–49. https://doi.org/10.1007/s10649-020-09962-3.
[57] Chang Y, Brickman P. When group work doesn't work: insights from students. CBE—Life Sci Educ 2018;17(3):ar52. https://doi.org/10.1187/cbe.17-09-0199.
[58] Hall D, Buzwell S. The problem of free-riding in group projects: looking beyond social loafing as reason for non-contribution. Act Learn High Educ 2013;14(1):37–49. https://doi.org/10.1177/1469787412467123.
[59] McKay J, Sridharan B. Student perceptions of collaborative group work (CGW) in higher education. Stud High Educ 2024;49(2):221–34. https://doi.org/10.1080/03075079.2023.2227677.
[60] Kennedy GJ, editor. Peer-assessment in group projects: is it worth it? In Vol. of 42 Proceedings of the 7th Australasian Conference on Computing Education, edited by Alison Young and Denise Tolhurst, 59. Newcastle: Australian Computer Society Inc; 2005. p. 65.
[61] Planas-Lladó A, Feliu L, Arbat G, Pujol J, Suñol JJ, Castro F, Martí C. An analysis of teamwork based on self and peer evaluation in higher education. Assess Eval High Educ 2021;46(2):191–207. https://doi.org/10.1080/02602938.2020.1763254.
[62] Holland D. Symbolic worlds in time/spaces of practice: identities and transformations. In: Wagoner B, editor. Symbolic transformation: the mind in movement through culture and society. New York, NY: Routledge/Taylor & Francis Group; 2010. p. 269–83.
[63] Holland D, Lachicotte W, Skinner D, Cain C. Identity and agency in cultural worlds. Cambridge, MA: Harvard University Press; 1998.
[64] Koretsky M, Nolen SB, Gilbuena DM, Tierney G, Volet S. Productively engaging student teams in engineering: The interplay between doing and thinking. Madrid, Spain: Frontiers in Education 2014; 2014. October 22-25.
[65] Nolen SB, Michor EL, Koretsky MD. Engineers, figuring it out: collaborative learning in cultural worlds. J Engin Edu 2024;113(1):164–94. https://doi.org/10.1002/jee.20576.
[66] Hall R, Jurow AS. Changing concepts in activity: descriptive and design studies of consequential learning in conceptual practices. Educ Psychol 2015;50(3):173–89. https://doi.org/10.1080/00461520.2015.1075403.
[67] Nolen SB, Wetzstein L, Goodell A. Designing material tools to mediate disciiplinary engagement in environmental science. Cogn Instr 2020;38(2):179–223. https://doi.org/10.1080/07370008.2020.1718677.
[68] Derry SJ, Pea RD, Barron B, Engle RA, Erickson F, Goldman R, Hall R, Koschmann T, Lemke JL, Sherin MG. Conducting video research in the learning sciences: guidance on selection, analysis, technology, and ethics. J Learn Sci 2010;19(1):3–53. https://doi.org/10.1080/10508400903452884.
[69] Jordan B, Henderson A. Interaction analysis: foundations and practice. J Learn Sci 1995;4(1):39–103. https://doi.org/10.1207/s15327809jls0401_2.
[70] Koretsky M, Falconer JL, Brooks BJ, Gilbuena DM, Silverstein D, Smith CL, Miletic M. The AICHE "concept warehouse": a web-based tool to promote concept-based instruction. Adv Engin Edu 2014;4(1). http://files.eric.ed.gov/fulltext/EJ1076138.pdf.
[71] Koretsky MD. Re-flipping in the remote classroom: the surprising uptake of video-recorded worked examples. J Chemic Edu 2020;97(9):2754–9. https://doi.org/10.1021/acs.jchemed.0c00711.
[72] Gavitte SB, Koretsky MD, Nason JA. Connecting affordances of physical and virtual laboratory modes to engineering epistemic practices. J Comp Higher Edu 2025;37(1):442–76. https://doi.org/10.1007/s12528-024-09403-7.
[73] Iiskala T, Volet S, Jones C, Koretsky M, Vauras M. Significance of forms and foci of metacognitive regulation in collaborative science learning of less and more successful outcome groups in diverse contexts. Instruc Sci 2021;49(5):687–718. https://doi.org/10.1007/s11251-021-09558-1.
[74] Koretsky MD, Amatore D, Barnes C, Kimura S. Enhancement of student learning in experimental design using a virtual laboratory [Article]. IEEE Trans Edu 2008;51(1):76–85. https://doi.org/10.1109/te.2007.906894.
[75] Koretsky MD, Kelly C, Gummer E. Student perceptions of learning in the laboratory: comparison of industrially situated virtual laboratories to capstone physical laboratories [Article]. J Engin Edu 2011;100(3):540–73. https://doi.org/10.1002/j.2168-9830.2011.tb00026.x.
[76] Koretsky MD. An interactive virtual laboratory addressing student difficulty in differentiating between chemical reaction kinetics and equilibrium. Comp Applic Engin Edu 2020;28(1):105–16. https://doi.org/10.1002/cae.22178.
[77] Bowen AS, Reid DR, Koretsky M. Development of interactive virtual laboratories to help students learn difficult concepts in thermodynamics. 2014 ASEE Annual Conference & Exposition. 2014. Indianapolis, IN.
[78] Friedrichsen DM, Smith C, Koretsky MD. Propagation from the start: the spread of a concept-based instructional tool. Edu Tech Res Devel 2017;65(1):177–202. https://doi.org/10.1007/s11423-016-9473-2.
[79] Hammer D, Berland LK. Confusing claims for data: a critique of common practices for presenting qualitative research on learning. J Learn Sci 2014;23(1):37–46. https://doi.org/10.1080/10508406.2013.802652.
[80] Walther J, Sochacka NW, Kellam NN. Quality in interpretive engineering education research: reflections on an example study. J Eng Educ 2013;102(4):626–59. https://doi.org/10.1002/jee.20029.
[81] Gilbuena DM, Sherrett BU, Gummer ES, Champagne AB, Koretsky MD. Feedback on professional skills as enculturation into communities of practice. J Eng Educ 2015;104(1):7–34. https://doi.org/10.1002/jee.20061.
[82] Korolija N, Linell P. Episodes: coding and analyzing coherence in multiparty conversation. Linguistics 1996;34(4):799–832. https://doi.org/10.1515/ling.1996.34.4.799.
[83] Van Dijk TA. Episodes as units of discourse analysis. In: Tannen D, editor. Analyzing discourse: text and talk. Washington, D.C: Georgetown University Press; 1982. p. 177–95.
[84] Glaser BG. The constant comparative method of qualitative analysis. Soc Probl 1965;12(4):436–45. https://doi.org/10.1525/sp.1965.12.4.03a00070.
[85] Pickering A. The mangle of practice: time, agency, and science. Chicago: IL: University of Chicago Press; 2010.
[86] Guzzomi AL, Male SA, Miller K. Students' responses to authentic assessment designed to develop commitment to performing at their best. European J Eng Educ 2017;42(3):219–40. https://doi.org/10.1080/03043797.2015.1121465.
[87] Chan RY. Understanding the purpose of higher education: an analysis of the economic and social benefits for completing a college degree. J Educ Policy Plan Adm 2016;6(5):1–40.
[88] Kaldaras L, Haudek K, Krajcik J. Employing automatic analysis tools aligned to learning progressions to assess knowledge application and support learning in STEM. Int J STEM Educ 2024;11(1):57. https://doi.org/10.1186/s40594-024-00516-0.
[89] Xia Q, Weng X, Ouyang F, Lin TJ, Chiu TK. A scoping review on how generative artificial intelligence transforms assessment in higher education. Int J Educ Technol High Educ 2024;21(1):40. https://doi.org/10.1186/s41239-024-00468-z.
[90] Hirshfield LJ, Koretsky MD. Cultivating creative thinking in engineering student teams: can a computer-mediated virtual laboratory help? J Comp Assist Learning 2021;37(2):587–601. https://doi.org/10.1111/jcal.12509.
[91] Nam BH, Bai Q. ChatGPT and its ethical implications for STEM research and higher education: a media discourse analysis. Int J STEM Educ 2023;10(1):66. https://doi.org/10.1186/s40594-023-00452-5.