

Priming the design process to improve student learning

*Nathan Mentzer, Abdul Moeed Asad, Wonki Lee,
Andrew Jackson, & Scott Bartholomew*

Abstract

Design thinking is critical in engineering and a fundamental for secondary students in introductory technology and engineering courses. The goal of this investigation was to measure impacts of a simple primer called Learning by Evaluating on students' design process. Students in introductory high school design courses were introduced to the primer as part of a treatment group. A comparison group did not receive the primer. Design documentation was collected in the form of students' design journals which were evaluated using a rubric. Results showed no significant differences in design journal scores between the group that received the treatment and the group that did not. While the primer did not result in measurable differences in design documentation, an interesting discovery emerged during the analysis. Design documentation, regardless of group, was not well aligned with standard practice. This discovery may warrant further investigation into how design documentation, a key design practice, is being implemented in classrooms and what kinds of support teachers may need to implement design with fidelity.

Key words: Comparative Judgement, Learning by Evaluating, design journal, design process documentation, design thinking

Mentzer, N., Asad, A., Lee, W., Jackson, A., & Bartholomew, S., (2026). Priming the Design Process to Improve Student Learning. *38*(1), 12-46.
<https://doi.org/10.21061/jte.v38i1.a.2>

Introduction

The need for high-quality STEM education has been well-documented by scholars and influential institutions, who argue it is essential for economic prosperity and an informed citizenry in these technologically-rich times (Jang, 2016; National Research Council, 2011; National Science Board, 2023). At an individual level, a strong STEM education foundation is essential for effective, evidence-based engagement in public discourse on critical scientific and technological issues (National Research Council, 2012). Beyond civic engagement, a STEM education sets students up for individual economic mobility and the nation for broader prosperity. According to the U.S. Bureau of Labor Statistics, STEM occupations are projected to grow at nearly three times the rate of other fields and offer higher median annual wages than that of non-STEM workers (Watson, 2025). At a national level, a STEM-trained workforce is crucial for maintaining a global edge in innovation (Jang, 2016), powering the

economic engine of a nation. Given these stakes, early and effective K-12 education is critical because research shows that a majority of students who pursue STEM careers make that decision during their secondary school years (Maltese & Tai, 2011).

The pressing need to improve STEM education requires the investigation of effective pedagogical strategies to better prepare students for an increasingly complex technological world. Accordingly, educators have moved toward teaching STEM education through a more cohesive, integrated approach that situates learning in authentic, applied contexts (Kelley & Knowles, 2016). Engineering design has emerged as a powerful catalyst for this integration, providing a natural framework for building connections among the STEM disciplines and earning a central place in national education standards (NGSS Lead States, 2013; *Standards for Technological and Engineering Literacy (STEL)* (ITEEA, 2020). However, the effective implementation of engineering design in the classroom remains a significant challenge. A consistent finding in the literature is that novice students struggle to engage in the design process as experts do, often rushing linearly through tasks and holding fundamental misconceptions about the nature of design (Crismond & Adams, 2012a; Mentzer et al., 2015). These struggles relate to both cognitive hurdles, such as underdeveloped problem-solving strategies (Crismond & Adams, 2012a), and affective challenges, such as a lack of self-efficacy in the face of the uncertainty inherent in design tasks (Cash & Kreye, 2018; Osmond & Tovey, 2015).

In response to the challenge of improving STEM education pedagogy by better supporting students' engagement with engineering design, this study investigates a pedagogical primer developed as part of a larger National Science Foundation funded project. This primer, known as Learning by Evaluating (LbE), engages students as judges of pairs of previous student work before engaging in similar work themselves as part of their current project. This design primer has potential to support novice designers to critically think about the nuances of what constitutes a good example of design work, thus actively engaging students in thinking about the nature of design as an intentional, thoughtful process that begins with a well-defined problem. In reviewing peer work, students may have a heightened sense of certainty that they too can be successful, which is especially critical for novice designers in open-ended ill-defined design tasks. With its foundation in cognitive apprenticeship and social constructivism, LbE has shown success in fostering student recognition of good design work and student design reasoning. Focusing specifically on secondary education, the current study builds on prior LbE research at both the middle school (Bartholomew, Strimel, et al., 2019) and university levels (Bartholomew et al., 2018; Bartholomew, Zhang, et al., 2019; Buckley et al., 2022; Seery et al., 2019; Seery & Canty, 2017).

The goal of this investigation was to empirically measure the effect of the LbE intervention on student learning and achievement within high school

engineering design classrooms. To that end, this study investigated the following research question: To what extent does the Learning by Evaluating (LbE) intervention impact student achievement in engineering design compared to standard classroom practice? This paper will review the relevant literature, present the research method and findings of this study, and discuss the implications of the results for technology and engineering education.

Literature Review

The Role of Engineering Design in STEM Education

The engineering design process is a complex cognitive activity including systematic and iterative activities to produce an output that satisfies requirements set by a client (Dym et al., 2005). The engineering design process stands as the central integrator of STEM education programs, with its significance recognized in integrated STEM academic frameworks (Kelley & Knowles, 2016), standards for technology and engineering (*Standards for Technological and Engineering Literacy (STEL)*, (ITEEA, 2020) and science education (NGSS Lead States, 2013) in the United States. The NGSS framework, a popular science education framework now adopted by 44 states (Next Gen Science, n.d.), identified engineering design across all three of its core dimensions, elevating engineering design as a fundamental element of STEM education (Moore et al., 2015). The emphasis on engineering design is reinforced by a pedagogical shift that moves from rote memorization of facts to a more process-oriented and problem-based approach to learning (NGSS Lead States, 2013). Beyond merely introducing students to subject knowledge, this movement signified a commitment to fostering essential 21st-century competencies, including systems thinking, creativity and collaboration because they are inherently practiced within the engineering design process (Lucas & Hanson, 2016; National Academy of Engineering & National Research Council, 2009). Thus, engaging in engineering design practices enables students to navigate an increasingly complex technological world.

The engineering design community has theorized that there are many classroom benefits attributed to engineering design, and particularly for the potential of integrated STEM education to improve student engagement in the pure sciences, increase overall learning and achievement, increase interest in the engineering profession, and improve technology literacy (Hill, 2006; Gattie & Wicklein, 2007; Feder et al., 2009). In the decades following the work of these scholars, there has been significant literature that has followed up on these claims (Sneider & Ravel, 2021) which has largely substantiated that the integrated teaching methods are more effective than traditional, siloed approaches. Sneider and Ravel (2021), in their systematic literature review, found that engineering design methods show their effectiveness with students self-reporting increased engagement in STEM education and better results in standardized tests of science and mathematics. As students move through the

iterative engineering design process, they build vital competencies for fostering innovation in professional engineering practice, referred to as the engineering habits of mind in the literature (Lucas & Hanson, 2016). These habits of mind include skills like system thinking, creative problem solving, collaboration and communication.

The Design Journal as a Tool for Authentic Assessment

Engineering design journals are used in engineering professional practice where, according to Kelley (2011), meticulous documentation is paramount. Due to the open-ended hands-on problem-solving emphasis of engineering design pedagogical practice in K-12 technology and engineering education, these journals are uniquely positioned instruments for authentic assessment in the classroom. They allow educators to evaluate the design process itself, not just the final outcome.

Much like professional engineers, students using design journals are tasked with chronologically recording their work. This documentation provides a comprehensive record of the design journey, encompassing background research, initial sketches, critical decisions, prototype evaluations, personal reflections, and iterative modifications (Kelley, 2011). To further mirror authentic engineering practices, students are often encouraged to follow professional conventions such as titling, signing, and dating each entry. The journal, therefore, becomes a compendium that prioritizes and makes visible the entire design process.

Building on this emulation of professional practice, scholars argue that design journals offer several distinct benefits for K-12 students. They are thought to foster metacognitive reflection, provide a rich source of data for assessment by making student cognitive processes visible, and instill good professional habits, such as refining communication skills and protecting intellectual property (Asunda & Hill, 2007; Hill, 2006; Kelley, 2011).

Current Challenges in Engineering Design Thinking Practices

While integrating subjects through engineering design holds immense promise for integrated STEM education, its classroom effectiveness may be contingent on proper implementation. To understand secondary student design thinking, Mentzer et al. (2015) compared the design processes of high school students with those of experts. Using an established protocol where participants designed a playground for a fictitious neighborhood, the researchers revealed that students tended to rush linearly through the critical early phases of problem-scoping and ideation. Experts, in contrast, dedicated substantially more time to these activities. This finding is consistent with a larger body of work by Atman and her colleagues (Atman et al., 2005, 2007), who used the same protocol to identify similar patterns between novice and expert designers at the university level. This tendency to shortcut the design process is a key challenge for

educators, and the literature suggests two primary explanations for it: one cognitive, rooted in student misconceptions about design, and the other affective, related to a lack of self-efficacy.

The cognitive explanation posits that students rush because they hold fundamental misconceptions about the nature of design itself. Design is not simply a linear procedure but a distinct way of knowing and reasoning (Cross, 2006). As Mentzer et al. (2015) showed, high school students have not developed the habits or expertise to operate at the cognitive level of experts. For example, according to Mentzer (2014), high school seniors' information gathering focused on solution-oriented information (e.g., information on material costs), instead of problem-focused information (e.g., community concerns) as experts' might. Mentzer's (2014) findings are consistent with Crismond and Adams' (2012b) influential *Informed Design Learning and Teaching Matrix* that highlighted nine key ways novices misconstrue design practices, which are detrimental to the final output. The misconceptions identified by Crismond and Adams span the entirety of the design process, from framing problems to reflecting on the final output. For example, like in Mentzer et al.'s (2015) study, novice designers are known to skip early research in favour of moving directly towards the solution, leading to missed opportunities in understanding the problem and subsequently missed potential in the quality of the designed solution. It is important to find effective ways to address students' misconceptions because misconceptions can be persistent and interfere with learning (Smith III et al., 1994).

A second potential explanation for novices' struggles is affective, centering on the concept of self-efficacy – an individual's belief in their ability to successfully execute the actions required for a specific task (Bandura, 1977). When students have low self-efficacy regarding design, they are less likely to persist through challenges and more likely to avoid behaviors they perceive as risky, such as open-ended creative exploration. This lack of confidence in a design context appears to stem from two interconnected sources: students' perceptions of their own abilities and the inherent uncertainty of the design process itself.

First, a student's self-concept can directly influence their design behavior. According to a study by Fila et al. (2014), college engineering students did not see themselves as creative, seeing a creative identity as something separate from an engineering identity. This led them to avoid activities they considered to be creative and gravitate towards routine and already existing solutions. Second, design is famously an uncertain, ill-defined process (Cash & Kreye, 2018; Osmond & Tovey, 2015; Tracey & Hutchinson, 2016). This ambiguity can lead to feelings of anxiety, and manifest as premature narrowing of creative exploration and rushing towards a solution in the engineering design process. Because navigating this uncertainty is so fundamental to the discipline, some

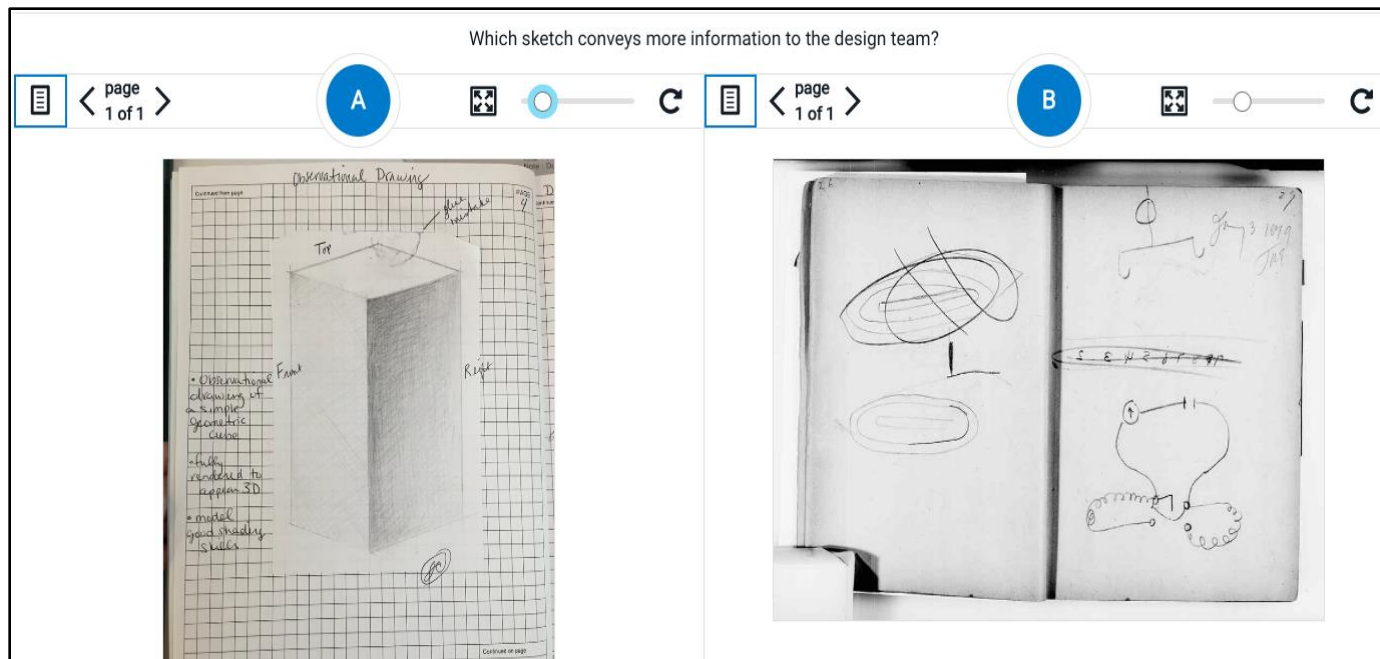
educators consider it a *threshold concept* that students must grasp to truly understand and feel efficacious in design (Osmond and Tovey, 2015).

Learning by Evaluating as a Technique with Potential to Improve Student Learning Outcomes

To address these challenges and improve student learning outcomes in engineering design, pedagogical strategies are needed that build student cognitive skills and sense of self-efficacy effectively. Learning by Evaluating (LbE) is a promising pedagogical technique hypothesized to achieve these goals because it is situated at the intersection of two powerful educational approaches: the use of exemplars and cognitive apprenticeship. This study, as part of a larger investigation into LbE, sought to understand its impact on student learning in an engineering design context (Bartholomew, Mentzer, & Jackson, 2022; Mentzer et al., 2024; Thorne et al., 2024).

LbE is a pedagogical technique used to prime students for learning experiences through several phases including conducting pairwise discrimination between exemplars based on a predetermined criterion (Figure 1). Comparison prompts are provided to the student by the instructor based on the learning objectives of the teacher (e.g., “which is a better problem statement?”) and, following a series of individual comparisons, a classwide discussion reinforces students’ understanding of quality design and equips them for project work. For examples of LbE sessions and findings related to students thinking as they engage in LbE, see Lee et al., 2024, 2025.

LbE works on the principle of comparative judgement to reduce the cognitive demands for students, allowing them to view multiple examples quickly. Comparative judgement is predicated on the fact that humans find it easier and are often more reliable when they are comparing two items side-by-side than when attempting to assign an absolute judgement to a single item in isolation (Bartholomew, Mentzer, & Jackson, 2022; Pollitt, 2012). LbE has shown to be beneficial in terms of elevating student learning for many disciplines including engineering design, English and business (Bartholomew, Mentzer, Jones et al., 2022).

Figure 1*Illustration of Pairwise Comparison in LbE Using RM Compare*

Note. The RM Compare interface displays the comparison prompt at the top ("Which sketch conveys more information to the design team?") followed by each item. Once an item is selected, the judge is prompted to enter their rationale.

LbE Exemplars for Building Understanding

Students in an LbE experience view exemplars that are usually snippets of varying quality from iterations of similar, previously conducted projects. In an engineering design context, exemplars are usually items from student design portfolios or pictures of prototypes. Exemplars work in two major ways: (1) By viewing quality examples, students are able to witness for themselves what successfully fulfilling criteria looks like in practice (Sadler, 2002), better aligning expectations with that of the instructor and reducing the ambiguity of assessment criteria; (2), by viewing poor quality examples, they are able to see that their own initial ideas are not as far off as they might have feared, building the confidence to begin the task (Canty, 2012; Hawe et al., 2021).

A common concern of using exemplars is that they can inhibit creativity or encourage students to copy from what they observe. However, research suggests that these concerns can be mitigated by careful pedagogical design. Some strategies include having students evaluate a range of ways to approach the problem, using exemplars that are good but not perfect, and using exemplars that are analogous but not identical (To et al., 2022). In addition, the open-ended, ill-defined nature of design problems is an advantage here because there is not one correct answer, there are many potential solutions. Indeed, given the variety of criteria that might relate to a design context, different solutions may excel in different ways (e.g., an idea that is lightweight vs. one that is durable vs. one that is economical). Making this point clear by discussing tradeoffs in the context of prior work provides students and their teacher a chance to clarify design value structures before embarking on their project.

LbE as a Form of Cognitive Apprenticeship

Cognitive apprenticeship can help students develop the cognitive skills necessary to engage in the engineering design process (Jackson & Strimel 2018). According to cognitive apprenticeship theory, developed by Collins et al. (1991), students gain expertise by observing and reflecting on the knowledge of others, including experts and peers (Collins, n.d.). The emphasis of the technique is to make thinking visible, allowing learners to access the tacit knowledge that guides expert performance. This model is particularly pertinent to open-ended disciplines like engineering design where problem solving is complex and solutions are not predetermined (Collins et al., 1991).

The LbE process can be framed as a powerful form of cognitive apprenticeship. The model of cognitive apprenticeship outlines six key teaching methods: modeling, coaching, scaffolding, articulation, reflection, and exploration. LbE directly engages three of these core methods (Bartholomew et al., 2024; Thorne et al., 2024). When students examine a pair of exemplars, they are going through the *modeling* phase of cognitive apprenticeship by observing how others perform a task. The act of making a judgment forces them to engage in *reflection* by having them compare the two artifacts against their own internal

understanding of quality, improving their personal mental models surrounding the task. Then, as they write the reasoning behind their choice, students engage in *articulation*. This act helps make their tacit knowledge explicit, which can lead to deeper understanding. Finally, students have the opportunity to test their ideas in a teacher-led debrief about why certain exemplars were ranked higher than others. This process externalizes their own thinking, allowing it to be examined, challenged, and refined through dialogue with peers and the instructor.

Method

Research Design

This study is part of a larger study on Learning by Evaluating involving multiple schools and teachers. However, the present study examined artifacts from only two teachers in the district. We selected these two teachers because we found that teachers in these classrooms were most consistent in expecting students to document design process work in their journals, allowing us the opportunity to analyze the quality of students' design thinking.

Our research followed a quasi-experimental nonequivalent treatment and control group design in order to make causal inferences about changes to student learning associated with the LbE experiences. Each technology and engineering education (TEE) instructor taught at least two sections of an introductory TEE course called Foundations of Technology. In our study, we randomly assigned treatment and control conditions to each teachers' classroom. Therefore, each instructor contributed both treatment and control students for this study. Instructional procedures which differed between these conditions are described later. To mitigate the potential for selection bias inherent to quasi-experimental nonequivalent treatment and control study designs, as described later we also measured our sample students' performance at the start and end of the term to increase internal validity (Shadish et al., 2001).

Curriculum Context

The International Technology and Engineering Educators Association (ITEEA) offers nationally and internationally disseminated curriculum and teacher professional learning opportunities. The association's introductory high school course is called Foundations of Technology (Gensemer & Caron, 2010). The study took place within this course. Foundations of Technology is centered on a series of iterative design cycles; in each, students work individually or in teams to design, build, test, refine, and finally present a solution to a design challenge. For example, one unit challenges students to improve a community park while learning computer-aided drafting (CAD) software. Instruction begins with students learning about CAD features and commands. Then, the class identifies a community park to analyse and improve. Students document their work in an engineering design journal (e.g., researching the problem, imagining

solutions, and so on) including design drawings and a write-up about the park improvements.

In this course, students are expected to keep a design journal documenting their design process. While design solutions, often products, are the focal culminating result of student work, the course learning objectives and content foster development of students' ability to work through a structured, iterative design process. Students' design journals are not only valuable for students to document their work for future reference and reflection, but as windows into student thinking processes for assessment purposes.

Procedure

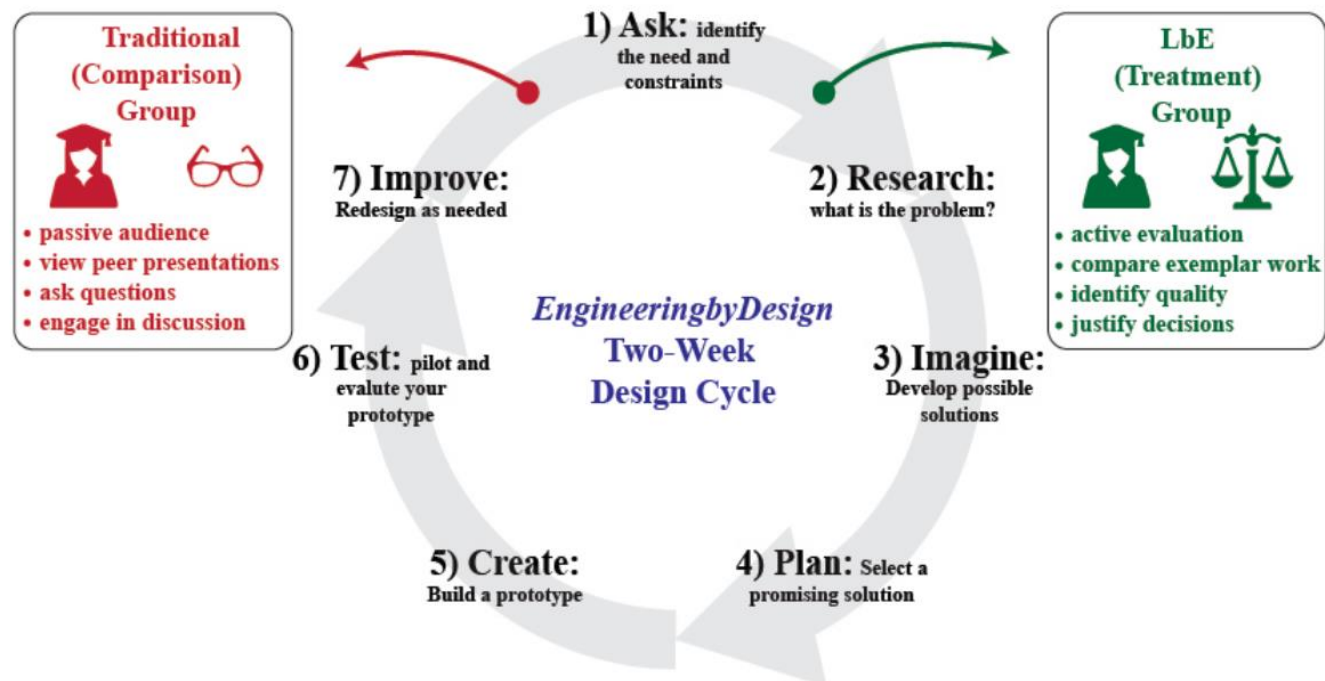
The teachers worked closely with the research team to foster fidelity of implementation and researcher awareness of curriculum in situ. Teachers were engaged in professional development on LbE annually in the form of a one day in person group meeting. In addition, teachers met with the research team at field based conferences, during observation visits and virtual visits throughout the year. The research team observed classrooms approximately monthly, engaging in the consent process as well as noting teacher strategies and offering support.

Control Group

The control group followed a traditional design cycle as shown in Figure 2 with a culminating report out, and often a peer presentation experience, as the final activity of the design cycle. These commonly took the form of a classwide presentation by each design team about their process, or a design report submitted to the teacher. These peer critique presentations gave students an opportunity to see others' work and engage in discussion around the process and solutions as a culmination of the design challenge or unit of instruction.

Figure 2

Illustration of the Differences Between Control and Treatment Curriculum Design



Treatment Group

In contrast, the treatment group participated in Learning by Evaluating (LbE) experiences at the start of a design cycle as shown in Figure 2 to prime students for the learning experience ahead. Not only did teachers use LbE at the start of the project, they also used it at the start of specific phases of the design cycle, such as problem definition or brainstorming. The LbE experience can be completed quickly (e.g., 10-15 minutes) and as alluded to in the literature review, is structured into three distinct phases:

Phase 1: Introduction and Orientation (Whole Group)

In the initial phase the teacher introduced the LbE session and oriented students to its learning goals. Sessions were created to fill in important gaps in student understanding as identified by the instructor and were aligned with the teacher's learning objectives. These objectives informed selection of exemplars and the creation of a specific comparison prompt that guided student judgments. For instance, if the objective was to improve prototyping skills, the instructor used images of prototypes on a similar project and constructed a comparison statement which might be as simple as, "*Which solution is better? Why?*" For a more targeted focus, such as structural integrity, the prompt could be refined to, "*Which solution can hold the most amount of weight? Why?*" During the introduction, the teacher often modeled the evaluation process by thinking aloud while making a comparison, or by leading a classroom discussion on the goals of the session (e.g., "*what are we looking for when judging structural integrity?*"). This modeling aimed to build students' capacity for critical evaluation and articulated reasoning (Jackson et al., 2022).

Phase 2: Individual Pairwise Comparison (Individually done on RM Compare)

In the second phase, students individually engaged in a pairwise comparison activity using the comparative judgement software *RM Compare* (RM plc, 2025). *RM Compare* software offered a secure online comparison system that harvested student activity data (individual student selection choices, comments, and class rank order). Students were presented with pairs of examples and tasked with determining which was superior based on the provided comparison prompt. A key component of this phase was the requirement for students to articulate the reasoning behind their decisions, thereby fostering critical thinking and engagement. As students in the class made selections, the *RM Compare* platform algorithmically generated a rank-ordered list of the examples based on the collective judgments.

Phase 3: Debrief and Reflection (Whole Group)

The final phase consisted of a whole-group reflection session led by the teacher, designed to facilitate the sharing of student thinking and to converge on a

common understanding of success criteria. During this debrief, instructors had the opportunity to reinforce key concepts and address any emergent misconceptions. The rank-ordered list generated by RM Compare often served as a focal point for discussion. Teachers typically used the highest-ranked examples to explore the qualities that constituted excellence, and the lowest-ranked examples to clarify common pitfalls and areas for improvement. This reflective discussion provided students with an opportunity to refine their mental models of high-quality work in preparation for embarking the project (or next phase of the project).

Data Collection

Data for this study were collected during the 2023-24 and 2024-25 academic years. With IRB approval, all students enrolled in the foundational course were invited to participate. Assent was obtained from all participants, as well as parent/guardian consent. Students were given the option to provide demographic data, including their race, gender, class year, and any prior experience in engineering or technology courses. A full summary of participant demographics is provided in the next section. At the conclusion of each academic term, researchers visited the classrooms to scan the physical design journals of all consented students. Digital scans of the journals were immediately uploaded to a secure, password-protected cloud server, and redacted prior to analysis.

Participants

This study was conducted in partnership with a large, diverse urban school district in the south eastern part of the US. The district serves over 100,000 students across more than 125 schools with a student population that is predominantly from underrepresented minorities in STEM fields (63.9% Black, 16.7% Hispanic, 11.0%, White, 6.4% Asian/Pacific Islander). Both teachers involved in the study had prior professional development in TEE and the LbE approach, as part of the larger ongoing project, and had previously taught the introductory course. The student population within their classrooms reflected the diversity of the broader district (see Table 1). One hundred and eighty-eight students had the opportunity to participate in the study but securing student and parental consent was challenging. The overall consent rate was 36.2%, with 17 students in control group classes and 36 in treatment group classes.

Table 1*Participant Demographics by Experimental Group*

	Treatment (<i>n</i> = 36)		Control (<i>n</i> = 17)		Full Sample (<i>n</i> = 53)	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Gender						
Male	17	47.2	14	82.4	31	58.5
Female	15	41.7	3	17.6	18	34
Non-Binary	0	0	0	0	0	0
No response	4	11.1	0	0	4	7.5
School Year						
Freshman	19	52.8	6	35.3	25	47.2
Sophomore	8	22.2	7	41.2	15	28.3
Junior	1	2.8	2	11.8	3	5.7
Senior	4	11.1	2	11.8	6	11.3
No response	4	11.1	0	0	4	7.5
Race						
Hispanic or Latino	10	27.8	7	41.2	17	32.1
White	15	41.7	2	11.8	17	32.1
Asian	3	8.3	4	23.5	7	13.2
Black or African American	2	5.6	2	11.8	4	7.5
Other or unknown	0	0	2	11.8	2	3.8
More than one race	1	2.8	0	0	1	1.9
American Indian or Alaska Native	1	2.8	0	0	1	1.9
No response	4	11.1	0	0	4	7.5
Prior Technology Experience						
No	30	83.3	14	82.4	44	83
Yes	2	5.6	3	17.6	5	9.4
No response	4	11.1	0	0	4	7.5

Of the participating students, 58.5% self-identified as male, 34.3% as female, and the remaining 7.5% did not respond to the question. The control group consisted of substantially more male students than female students. As expected for the introductory course, participants were primarily underclassmen, with freshmen (47.2%) and sophomores (28.3%) comprising the majority. The remaining consisted of juniors (5.7%), seniors (11.3%), and those that did not respond (7.5%). The largest proportion of students identified as Hispanic or Latino and White with each comprising around a third of total students (32.1%). Asian students were the third largest group making up 13.2% of the study sample. Proportionally, the treatment group had four times as many White students while the control group had twice as many Asian students. About a tenth of students (9.4%) had experience with TEE courses in the past, but most indicated they were taking a TEE course for the first time (83.1%).

Data Analysis

Measurement Instrument

Student performance was measured using a rubric produced as part of a previously funded \$1.7M NSF DRK12 Grant (#1316601) project titled Engineering for All (EfA). The EfA design journal rubric was created to measure design practice standards that were rated by experts as *feasible*, *important*, and *clearly written* in the previous study (Lomask et al., 2018). In our study's context, the EfA rubric was in close alignment with the design process and expectations associated with the International Technology and Engineering Educators Association (ITEEA) Foundations of Technology course for classroom assessment. Therefore, we expected direct alignment between how students are evaluated in their courses and how their learning was measured in this research. The rubric used in this study for measuring student performance can be found on the EfA design project's website (<https://www.hofstra.edu/stem-research/engineering-for-all.html>).

The rubric consists of *nine criteria* that assess the engineering design process: (1a) *stating the design challenge*, (1b) *clarifying criteria*, (1c) *clarifying constraints*, (2) *research and investigation*, (3) *generating alternative solutions*, (4) *choosing and justifying the optimal solution*, (5) *developing a prototype*, (6) *testing and evaluating the design solution*, and (7) *redesigning the solution*. Each criterion was scored on a 3-point scale, which we modified to be a 4-point scale by including a zero level to explicitly account for students who did not attempt the criteria even though it was expected. The original scale was scored with a 1 where the performance was below expectation and "needed improvement", a 2 for "acceptable" work that was at expectation for the criteria being measured and a 3 for "really done well" such that the student had exceeded expectations. Each of the 9 criteria had additional context specific definitions for the scores 1, 2 and 3.

The primary outcome measure for this study was the overall project-level rubric score, calculated for each design journal by isolating a project, scoring it, and averaging the results of the scorable criteria. Although teachers for this study were chosen based on the general level of documentation in their classes, we found differences in which design documentation elements were required in each class. Therefore, when criteria have missing data, we chose to use available criteria when calculating the rubric score (not all 9 criteria). Average scores of available criteria were used to enable comparisons between pre- and post-measures and between control and treatment conditions. While project-level scores were used to test the main hypothesis, criterion-level averages are also reported in the findings section to illuminate student performance on each criterion and highlight what each criteria measured.

To obtain pre- and post-intervention scores for each student, we analyzed design journals from two separate projects selected based on included dates and chronological order. The pre-test score was derived from the first substantial project of the course which was either a pin design and a crane design. The post-test score was derived from the final substantial project which was either a fidget spinner or a bridge design. For our investigation, a *substantial project* was defined as one spanning at least two weeks, distinguishing it from shorter (one day), introductory exercises.

Interrater Consensus

To ensure calibration and reliability of scoring with the rubric, a multi-phase process was implemented. Two authors independently scored the same 4 design journals representing about 7.5% of the sample. Inter-rater reliability for the rubric scores was assessed using a two-way random-effects, absolute agreement, average-measures Intraclass Correlation Coefficient. The ICC was .57, $F(35, 35) = 2.43$, $p < .005$, 95% CI [.17, .78] which was considered moderate agreement. The researchers then discussed and negotiated any scoring discrepancies to consensus. Notes taken during this process led to an annotated rubric with documented details to clarify the operational definitions of each score category and to standardize the handling of atypical student journal entries. For example, an annotation was made for Criterion 7 (*Redesign the Solution*) to specify that qualitative data could be considered equivalent to *quantitative data* (specified in the original rubric) for a score of 3, which better aligned with teacher expectations of student work. A second round of independent scoring was conducted using 4 additional journals (7.5%) and, again, discrepancies were negotiated and documented. The ICC was .83, $F(35, 35) = 5.91$, $p < .001$, 95% CI [.67, .92] which was considered good agreement. At this point, the authors agreed the consistency between them was sufficient to have one author, blind to the treatment/control condition of each journal, code the entirety of the data set.

Findings

The findings are presented in two parts. We begin with the project-level analysis to investigate design documentation of students in the control group compared to those in the LbE treatment using gains from pre-to-post measures. Then, we present descriptive scores on each of the 9 rubric criteria to provide context and a more nuanced understanding of student design documentation.

Project-Level Analysis

To address our study's driving research regarding measurable learning gains related to LbE experiences, we conducted a repeated measures ANOVA using gain scores between the pre- and post-measures of student design documentation. Descriptive statistics show a modest increase in mean scores for both groups over the course of the study. The Treatment Group ($n=36$) saw their mean score increase from 2.20 to 2.34. The Control Group ($n=17$) saw their mean score increase from 1.97 to 2.23.

Prior to a repeated measures ANOVA, data were checked for normality using a Shapiro-Wilk test due to the small sample sizes. Pre- and post-scores were both found to deviate significantly from normal distribution. Therefore, we examined the Greenhouse-Geisser measure of Sphericity. The epsilon values were 1.000 indicating that Sphericity Assumed univariate test results were appropriate for interpretation of the repeated measures ANOVA tests. Effect size interpretations were based on Cohen (1988) partial η^2 values of small, medium, and large as .0099, .0588, and .1379, respectively.

The analysis revealed a statistically significant main effect of time indicating that students' scores improved significantly between the first measure and the last, $F(1, 51) = 5.635, p = .021$. The effect size, partial eta squared (η^2), was 0.099, indicating the measure of student learning showed a small effect. Though the effect was small, the significant gains align with the course goals and lend credibility to our measurement strategy as teachers were strategically targeting learning of the design process which was measurable with our data and instrument.

However, the analysis of the time-by-condition interaction effect was not statistically significant, $F(1, 51) = .466, p = .498$ with a negligible effect size of .009. This crucial result indicates that the change in scores over time for the LbE treatment group was not significantly different from the change in scores for the control group. Thus, the treatment did not have a measurable impact on learning.

In summary, while students' design journal scores improved, these findings show no statistical evidence that the Learning by Evaluating treatment was more effective than standard classroom practice in producing this improvement.

Criterion Level Results

While the overall statistical analysis showed no significant treatment effect, examining the results at the criterion level reveals a more detailed picture of

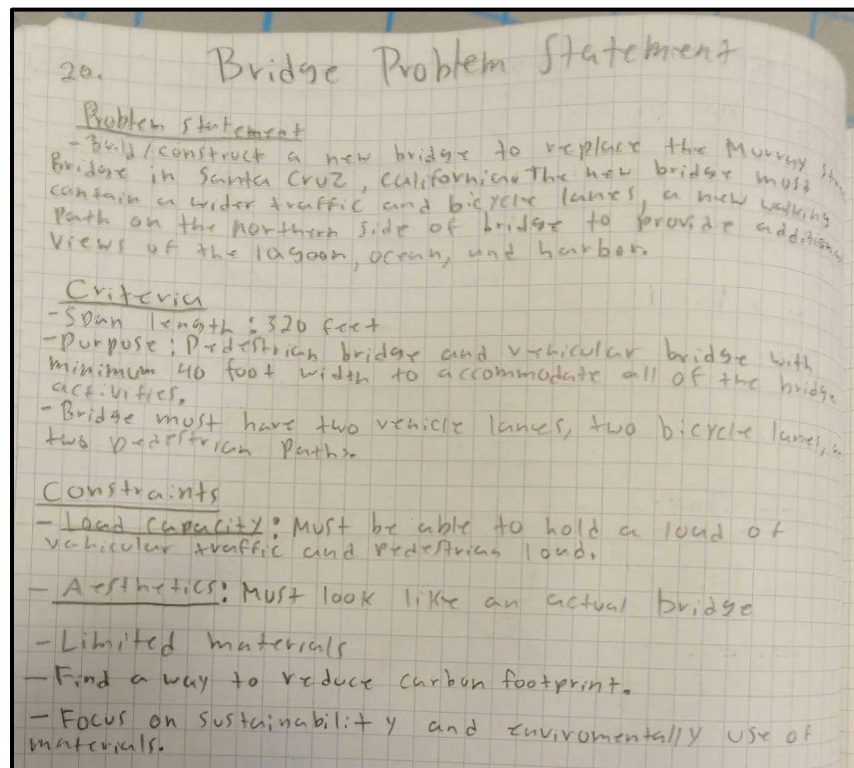
student documentation performance. The following section provides a descriptive breakdown of the pre- and post-test scores for each individual criterion, highlighting specific patterns of change and makes explicit our interpretation of the rubric.

Criterion 1a: Stating the Design Challenge

An example of the first several criteria is shown in Figure 3. For this criterion, the pre-score mean was 1.87 (n=53) and post-score was 1.81 (n=53), indicating very slight loss during the course. Most students received a score of 2 ("Acceptable") by restating the design challenge as it was provided by the instructor.

Figure 3

Example of Student Design Journal with Criterion 1a (Stating the Design Challenge), 1b (Clarifying Criteria) and 1c (Clarifying Constraints)



To achieve a score of 3, according to the rubric, students needed to frame the challenge in a way that also explained the underlying need for the work. Only four students, all of whom were in the post-test group, achieved this score. For example, one student who earned a 3 wrote:

Build/construct a new bridge to replace the Murry Street Bridge in Santa Cruz, California. The new bridge must contain wider traffic and bicycle lanes, a new walking path on the northern side of the bridge to provide additional views of the lagoon, ocean and harbor.

The inclusion of the rationale (the need) for the new features that went beyond what was provided by the instructor, “provide additional views of the lagoon, ocean and harbor”, differentiated the level of performance as “Really Well Done”.

Criterion 1b: Clarifying Criteria

A noticeable improvement was observed for Criterion 1b, with the mean score increasing from 1.58 ($n=53$) in the pre-test to 2.55 in the post-score ($n=53$). An observation while scoring was the frequent confusion between design criteria and constraints. For instance, one student listed *load capacity*, *aesthetics*, and *reducing carbon footprint* under the constraints section (see Figure 2), when these are all performance criteria. Conversely, the same student listed *span length* and specific requirements like *two vehicle lanes* under the *Criteria* section, when these actually function as constraints on the design.

For purposes of this research, we focused on the existence of criteria regardless of them being categorized as criteria or constraints. Despite this, many students in the pre-test still struggled to identify any criteria, leading to scores of 1 (*Needs Improvement*). To receive a score of 2 (*Acceptable*), according to the rubric, students were required to state two distinct criteria for their design. A score of 3 was awarded to students who identified multiple criteria. The notable increase in the post-test score indicates that over the course of the intervention, students became significantly more adept at identifying and articulating the desired attributes of their designs, even if the formal distinction between criteria and constraints remained a point of confusion.

Criterion 1c: Clarifying Constraints

For this criterion, students began with a high level of proficiency that improved slightly, with the mean score increasing from 2.30 ($n=53$) to 2.55 ($n=53$). As mentioned, students often confused constraints with criteria; however, our scoring protocol awarded credit for the existence of constraints regardless of how they were labeled.

To score a 2 (*Acceptable*), students only needed to identify one constraint. Despite this relatively low expectation, a small number of students scored a 0 for not attempting the criterion or a 1 for failing to identify any constraints. Yet, an interesting discovery on this criterion was a pronounced ceiling effect. The median score for both the pre-test and post-test was 3, indicating that the majority of students successfully identified multiple constraints and earned the highest possible score from the outset.

Criterion 2: Research and Investigation

The mean score for this criterion showed a slight increase, moving from 1.63 in the pre-test ($n=30$) to 1.70 in the post-test ($n=53$). The discrepancy in sample size was due to one instructor's students not documenting parts of their project, rendering their pre-test journals unscorable on this criterion. However, the broader inclusion of this documentation following the introductory course is indicative of increased sophistication with students' design thinking and documentation in general. Students' research and investigation for both teachers' projects included students entering a list of *research questions* in their design journals and then using internet searches to answer them with occasional citations documenting sources.

A notable finding from the data was the prescriptive nature of the student work. Journal entries were highly uniform within classes, suggesting that the format was heavily guided by the instructor. In one teacher's post-test project, for example, all students documented their research by posing exactly five questions to be answered. This templated approach meant that most students achieved a score of 2 (students completed the pre-design tasks), perhaps just because they followed the teacher's requirements.

The primary differentiator for scoring was the quality of the research questions. Students scored low on the rubric for including questions that were not directly useful for constructing their prototype. Questions like *What is [the product]?* were regarded in our scoring as unhelpful, as it targeted general information students should have already known and therefore scored a 1 indicating students had not completed the research as a pre-design task. In contrast, questions such as, *What makes [the product] function well?* was considered helpful for informing the design and was therefore regarded as *completing research as a pre-design task*. To achieve a score of 3, students needed to do additional research that informed their design solutions which was interpreted as going beyond the teachers' expectations.

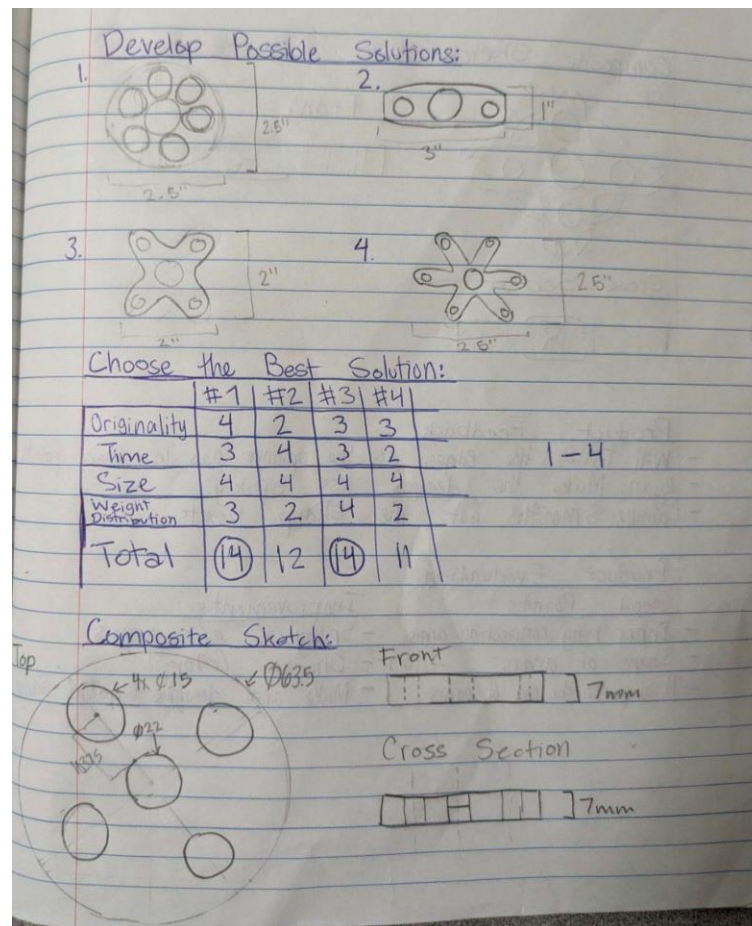
Criterion 3: Generate Alternative Solutions

The mean score for this criterion was high for both pre- and post-tests, showing a slight decrease from 2.97 ($n=30$) to 2.77 in the post-test ($n=53$). As with criteria 1c, we observed a strong ceiling effect, with most students achieving a score of 3. To earn this maximum score, students needed to show

evidence of an iterative process (Figures 4 and 5), by brainstorming multiple distinct solutions and then presenting a final sketch that differed from their initial ideas, therefore showing iteration. A score of 2 was given if students did not show clear evidence of this iterative problem-solving but did have a few ideas. A score of 1 was designated for students who only presented a single design without considering alternatives; notably, no student received this score in either the pre- or post-test.

Figure 4

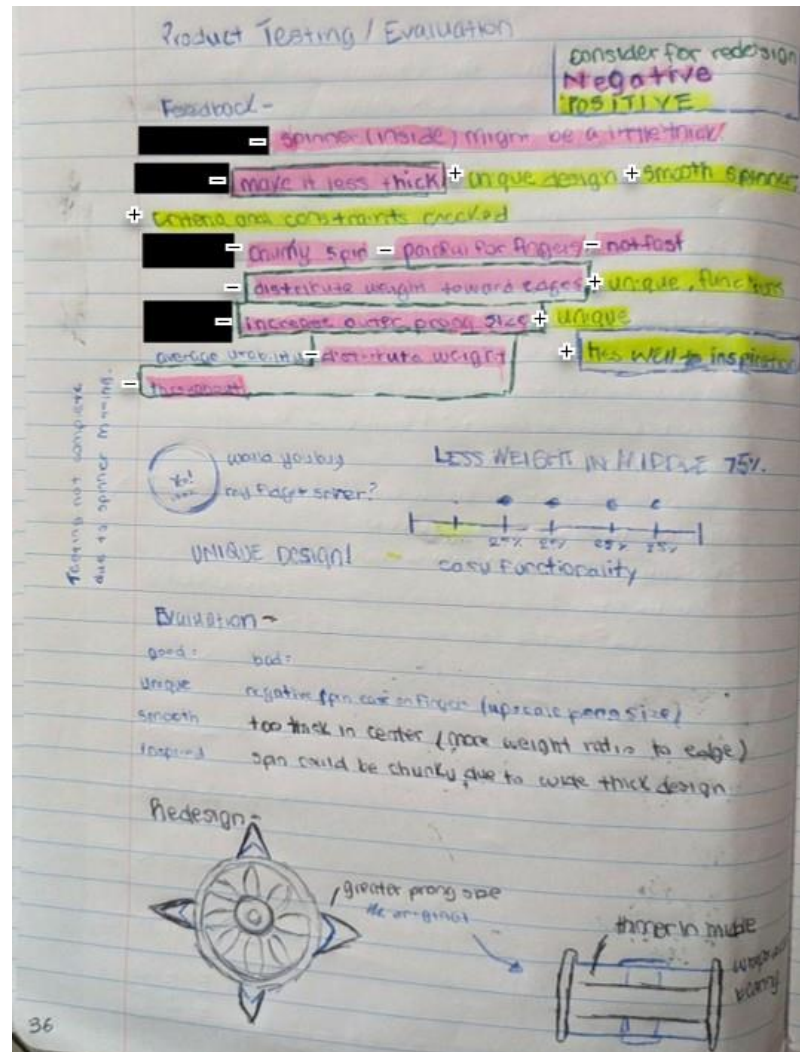
Example of Iterative Sketching in Design Documentation



Note: A student brainstormed 4 initial designs and created a different final sketch after evaluating all of their ideas earning them a 3.

Figure 5

Example of Iterative Sketching in Design Documentation



Note: A student collected feedback from their peers (names redacted) on their prototype and has highlighted them as “positive” (yellow) and “negative” (pink) using a marker followed by a redesign which differed from their initial ideas (not shown) resulting in a score of 3. To distinguish between peer feedback of student work in the black and white print copy, the authors have added (+) to indicate positive and (-) to indicate negative to the left of each item of feedback.

Criterion 4: Choose and Justify the Optimal Solution

The mean score for this criterion was also high, increasing from 2.33 in the pre-test ($n=30$) to 2.51 in the post-test ($n=53$). Performance levels on the criterion required a decision with justification for the highest score, while limited justification warranted a 2, and a lack of discussion earned a score of 1. Students who had no clear evidence of making a choice earned a 0. Surprisingly, scores on the criterion were bimodal, due to the common use of a decision matrix by students. This meant that students either had a matrix, which facilitated justifications and decisions, to earn a 3 or didn't have any evidence of making a decision and earned a 0.

As researchers, our evaluation of the quality of reasoning around the matrix was limited. The rubric only required the presence of quantitative measures. Further, the ambiguity of student-written labels would have made subjective quality judgments difficult, though perhaps deeper exploration of student decision processes would reveal differentiators in student learning. An example of a decision matrix can be seen in Figure 4.

Criterion 5: Develop a Prototype

A notable decrease was observed for this criterion, with the mean score dropping from 2.85 in the pre-test ($n=53$) to 2.36 ($n=30$) in the post-test. This decrease was primarily driven by a lack of documented evidence of prototyping in the post-test journals from one instructor's classroom.

The scoring for this criterion was largely based on completion. The rubric required a *representation* of the prototype (such as a final design sketch) for a score of 2, and evidence of a *working model* for a score of 3. As very few students included photographic evidence of their prototypes, we inferred that a working model had been created if the student included testing-related data from their prototype teaching in their journal. If testing data were absent, students with a final design sketch received a score of 2. The limitations of this inferential approach are elaborated further in the discussion section.

Criterion 6: Test and Evaluate the Design Solution

The mean scores for this criterion were high at the start of the term and end, showing a slight decrease from 2.58 in the pre-test ($n=53$) to 2.47 in the post-test ($n=30$). The high scores were attributable to the prescriptive nature of the testing procedures. In the projects we observed, all students appeared to perform the same pre-determined tests, likely assigned by the instructor. Consequently, students who documented the completion of this required test, which was considered a *fair test* by the rubric and analyzed these results to determine strengths and limitations, received a score of 3. Students who had no documented testing data, received a score of 0. Still, this bimodal scoring pattern produced a high overall average.

Criterion 7: Redesign the Solution

A noteworthy increase was observed for this criterion, with the mean score rising from 1.55 in the pre-test ($n=53$) to 2.13 in the post-test ($n=30$). This improvement reflects students' increased ability to justify their design modifications. To achieve a score of 3, students were required to articulate the reason for their redesign by explicitly referencing quantitative data from their testing phase (Criterion 6). To allow for variations in projects, we also interpreted sufficient qualitative feedback to serve as valid data for criterion (see Figure 3 for an example of a student that earned a 3). Students who proposed a redesign but did not connect it back to their testing results received a score of 2.

Discussion

The results of our study indicated that Learning by Evaluation (LbE) treatment did not produce a statistically significant change in student achievement. This finding is unexpected, particularly given the robust theoretical foundations of the intervention and the past research which has consistently demonstrated significant positive outcomes from LbE application (Bartholomew, Mentzer, Jones, 2022). This section will rationalize why this result is surprising by examining the established literature on the core components of LbE and then explore potential alternative explanations for our findings.

Our expectation that LbE would positively impact student achievement was grounded in well-established educational theories. The integration of exemplars into a curriculum, a key feature of LbE, has been theorized to enhance student achievement by fostering self-regulated learning (To et al., 2022). In a systematic study on the use of peer exemplars in technology and engineering education, Thorne et al. (2024) identified multiple benefits that lead to enhanced learning, including improved student focus, increased motivation to learn, the development of self-efficacy, and a higher standard for learning outcomes. Similarly, the principles of cognitive apprenticeship have been validated by over three decades of scholarship, which consistently demonstrate its success in improving student learning and achievement (Collins et al., 1991; Dennen & Burner, 2008; Wu et al., 2016).

The impetus for this study was the documented success of LbE in other educational contexts. LbE has proven to be an effective technique in higher education (Bartholomew, Mentzer, Jones, et al., 2022; Bartholomew, Zhang, et al., 2019) and in middle school (Bartholomew et al., 2018) for enhancing student learning. We, therefore, anticipated measurable gains in student learning and achievement at the high school level. However, a critical distinction between our study and prior research may be the scope of our measurement. In the higher education studies, for instance, learning improvements were measured on specific, targeted tasks. In contrast, our study attempted to detect improvements at a broader project level, using average design journal scores as the primary

metric. Therefore, a plausible explanation for non-significant results is that our approach masked more granular improvements (i.e., time management or the quality of entries in a decision matrix that teachers may have targeted with their LbE sessions). Students may have experienced gains on specific design activities, such as producing more creative ideas or more complex prototypes, but these nuanced changes were not captured by the overall project score or rubric criteria associated with the specific design activity.

Furthermore, a potential mismatch likely existed between the content of some LbE sessions and the criteria assessed by our rubric. Ideally, we would analyse the specific LbE sessions each instructor ran and relate those to performance gains on specific design criteria, but our data collection was unable to capture which sessions were run by which teachers during this design journal study. We discovered that instructors often refreshed older LbE sessions for new activities instead of creating new ones, which meant we could not systematically track the precise content or frequency of the sessions that were actually run using the teacher's history of sessions in RM Compare. Nonetheless, our broader engagement with the classrooms as part of a larger NSF project allowed us to observe instructors applying LbE to valuable skills like improving the quality of brainstorming or enhancing time management via Gantt charts. While these are important professional practices, they are not directly measured by EfA design journal rubric. This suggests that students could have been learning and improving in ways that our study was simply not designed to measure. We will explore the limitations of our instrument in greater detail in the subsequent section.

Limitations Associated with the Rubric Fit

A significant limitation of this study was the alignment between the general measurement instrument and the specific LbE treatment. The rubric was designed for teacher led classroom assessment, which we suspect would have been successful based on our use. However, as a research instrument, it proved to be a poor fit for the needs of this particular study, which was not evident until analysis was completed. The rubric design prioritized the completion of tasks (appropriate for classroom teacher assessment—its designed intent) over a detailed inspection of performance quality (needed for this research).

This difficulty in examining fine-grained levels of performance was manifest in several elements of the rubric design. The rubric had a 3-point scale (1 to 3), which we expanded to a 4 point scale (0 to 3). However, the instrument still likely lacked numerical granularity to capture certain improvements, based on the typical design performance of the student participants. For instance, under *Research and Investigation* (Criterion 2), the rubric evaluated if research had been conducted, but not the quality of the research or its impact on the design work. We speculate LbE may improve students' critical thinking which may emerge in ways that are demonstrated through more meaningful research as

opposed to the binary presence of research or not. Thus, while parts of the rubric valued completion, we hypothesize LbE might improve quality, not completion. This represents a significant limitation for our study, as it renders a crucial aspect of intellectual development—the increasing sophistication of student inquiry—largely invisible to the assessment.

Several aspects of classroom scaffolding may also have amplified the difficulties of using the rubric to evaluate performance within a research study. In several cases, teachers' typical classroom practices created an artificial ceiling when combined with the rubric focus on completion rather than quality. We hypothesize this was the case for *Stating the Design Challenge* (Criterion 1a).

In our analysis, we observed that students consistently earned a score of 2 by *restated [the problem statement] exactly as provided by the instructor*. To achieve a 3, the rubric required students to independently reframe the problem to explain the need for the project. Although the higher level does reflect increased sophistication with design documentation, if students were only instructed to perform the action that earned a score of 2 (i.e., restate exactly as provided), it would explain why students typically did not achieve the highest level. This same pattern could have occurred with the *Research and Investigate* criterion, where students may have precisely followed instructions (i.e. research exactly 5 questions) which corresponded only to a mid-level score on the rubric.

Conversely, other criteria were susceptible to a different kind of ceiling effect, where the requirements for the highest score were met with relative ease. For *Clarifying Constraints* (Criterion 1c) and *Generate Alternative Solutions* (Criterion 3), most students identified several design constraints (probably teacher provided) and had multiple ideas that differed from their final idea and therefore received the maximum score of 3. While this indicates strong performance, it also means the rubric could not differentiate between students who met the high standard and those who significantly surpassed it.

Limitations Associated with the Design Journals

This study is situated within a larger, five-year NSF-funded project investigating Learning by Evaluating. While the current analysis was designed to assess student learning based solely on the evidence available within engineering design journals, our understanding of the context is informed by our broader engagement with the participating classrooms. The journals, however, often failed to provide a comprehensive record of the students' design processes. This created a significant discrepancy between the activities we witnessed during our ongoing observational visits and how students' work was actually documented.

This issue of incomplete documentation was widespread. Of the nine instructors involved in the larger project for the academic years of 2023-24 and 2024-25, only two consistently required students to maintain their work in a single, coherent design journal. While these two teachers had the richest

journals, even amongst them, journals were often missing key entries of the design journey, contrary to the best practices outlined by Kelley (2011), which emphasize diligent, chronological recording of all design activities.

The evidence of this incomplete documentation was most apparent in the scoring for *Develop a Prototype* (Criterion 5). During our observational site visits—which were part of the larger grant's activities but not a formal data source for this particular study—we frequently saw students constructing both digital and physical prototypes. Yet, because this work was not consistently recorded in the paper based journals, many students did not receive a high design journal score for their prototyping work, due to a lack of documented evidence. Our direct observations allow us to surmise with confidence that most students did, in fact, complete the prototyping stage, but they failed to include the necessary evidence in their journals. Instead, they may have been evaluated by the teacher's observations or there may have been digital submissions to the school-wide LMS software that captured the process but were not entered into the design journal. Consequently, relying solely on the design journals, as dictated by this study's methodology, provided an incomplete and potentially misleading representation of student achievement.

Implications for TEE

Implications for Practitioners

The findings of this study offer several key implications for technology and engineering practitioners. Our analysis showed a significant improvement in students' overall design journal scores over time, affirming that valuable learning is occurring regardless of our LbE treatment. However, our findings also highlight opportunities to enhance this practice by addressing two key areas: (1) the need for improved, practice-based documentation strategies and (2) a fundamental tension between prescriptive instruction and the development of authentic engineering habits of mind.

Re-envisioning the Engineering Design Journal as a Practice-Based Tool

A central tenet of using engineering design journals is to make student reasoning visible, thereby enabling authentic assessment (Asunda & Hill, 2007; Kelley, 2011). Our finding that most instructors did not use journals this way suggests there is a significant gap between pedagogical theory and classroom reality. This gap presents an opportunity to consider documentation approaches that are practice-based. Such a practice-based approach might be a better reflection of classroom dynamics, where the work students complete is often digital, collaborative, and multimedia-rich. Rather than being a static repository, contemporary, practical journals could be a dynamic digital space where students embed videos of prototype tests, link to simulations, and collaborate on technical drawings.

The gap between design journals as authentic assessment and their practical use appears to be one of integration; when documentation is perceived as an onerous task separate from the *real* work of designing and building, students and teachers may be less likely to engage with it deeply (Paretti, 2008). Future work should therefore focus on making documentation a more authentically practice-based endeavour. A promising approach could be to co-develop new resources directly with instructors, ensuring that solutions are grounded in classroom realities. This collaborative work could include creating new practice-based implementation frameworks or exploration with digital platforms that better integrate documentation into the design process, framing it as a natural and valued part of the engineering workflow. With these blended modalities and platforms, transitioning the historically paper based design journal to a digital structure may expand the comprehensiveness and utility of engaging in design documentation.

On the other hand, absent such a platform, effectively evaluating student design documentation may require that information be combined from several sources. Our observations indicated that students were making sketches on paper and doing CAD work on their computers. Prototype testing was done in class but could have been documented with video or photographic evidence. Therefore, integration of student design documentation may require examining paper-based journals, computer and empirical testing, digital LMS submissions, and in person observation when presentations are employed for assessment.

Balancing Instructional Scaffolding with Authentic Engineering Habits of Mind

Beyond documentation, our analysis raises a fundamental question about how design is taught in secondary education. While the stated ideals of the field emphasize the cultivation of engineering habits of mind, such as systems thinking (Lucas & Hanson, 2016), our analysis revealed student work that was often highly uniform, implying that it was prescriptive. As examples, students were expected to restate the design problem exactly as provided by the instructor and many of the constraints and criteria were identical within classes. A prescriptive approach, while efficient, is in direct tension with the goal of fostering authentic problem-solving. This tension may point toward a pedagogical *sweet spot* where instructors provide structure by defining the boundaries of a problem space (e.g., specifying certain materials, software, or a target performance metric), yet leave the pathway to the solution genuinely open-ended. Such a strategy could provide necessary guidance while still entrusting students with intellectual agency.

A valuable direction for future research, therefore, would be to investigate learning progressions associated with secondary design education with specific attention to scaffolding novice designers sufficiently. For example, comparative studies could examine differences in student learning, creativity, and systems

thinking when taught in highly prescriptive environments versus more open-ended, yet constrained, ones. Furthermore, another line of inquiry could focus on developing resources or frameworks that help educators design effective learning experiences that balance necessary constraints with the freedom to innovate.

Implications for Researchers

Rather than concluding that Learning by Engineering (LbE) may be ineffective in measurably improving student learning in high school settings, we argue the primary implication of this study is methodological. We have aimed for transparency in our research process so that subsequent work can better navigate the complexities of measuring LbE's effects. Our findings suggest that for results to be clear, there must be a much tighter alignment between the specific goals of an LbE intervention and the constructs being measured. A review of available instruments to measure secondary introductory level design process documentation was conducted prior to this study and resulted in few options. Thus, future research could refine a valid and reliable research grade instrument for measuring design to fill this gap.

With a more rigorous instrument our study's next weakness was sample size. Future research should be situated in classrooms where design documentation is an expectation. While nine teachers participated in the larger study, and most teachers had what they called design journals, they typically expected students to use these as notebooks rather than design journals. Thus, upon close review, we discovered the design journal was more of a collection of lecture notes, including anything students were asked to write down in preparation for a test, draft research papers, or notes when a guest speaker was coming into the classroom. To complicate our limited teacher sample, we recommend future researchers work closely with the school community during the consent process as we struggled to secure parent consent. Few parents actively said no, but, rather, our struggle was simply getting a response.

Conclusion

This study sought to determine if Learning by Evaluating (LbE) could significantly improve student achievement in secondary school engineering design. Despite LbE's promising foundations, our primary finding was that the LbE intervention did not produce a statistically significant improvement in student design journal scores compared to standard classroom practice. We conclude that this outcome is likely not an indictment of LbE itself, but rather a reflection of the limitations of the measurement strategy used in this study. While the rubric used may be suitable for classroom assessment, it was general and focused on task completion over quality. Its inability to capture nuanced or qualitative growth may have masked student learning associated with our

treatment. Second, a mismatch between LbE activities and the assessment criteria also likely rendered genuine learning gains invisible to our analysis.

Therefore, the primary contribution of this work is a critical insight into the challenges of authentic assessment in K-12 engineering education. Our findings compel a shift in focus from merely implementing new teaching techniques to fundamentally rethinking how we measure complex learning. This points to two clear implications for practitioners: the need for better integrated, practice-based documentation that make student thinking truly visible, and a pedagogical move away from prescriptive instruction toward creating more open-ended, yet constrained learning experiences.

Funding Information

This material is based on work supported by the National Science Foundation under Grant Number 2101235.

Conflict of Interest Statement

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

This study was approved by the Purdue University IRB and administrators of participating school districts. Participating teachers, students and students' parents provided consent and assent as required by Purdue University IRB. All participants were informed about the study during the consent/assent process which included the intention to publish the study findings, ensuring no individuals could be recognized.

References

- Asunda, P., & Hill, R. (2007). Critical Features of Engineering Design in Technology Education. *Journal of STEM Teacher Education*, 44(1).
- Atman, C., Adams, R. S., Cardella, M., Turns, J., Mosborg, S., & Saleem, J. (2007). Engineering design processes: A comparison of students and expert practitioners. *Journal of Engineering Education*, 96(4), 359–379. <https://doi.org/10.1002/j.2168-9830.2007.tb00945.x>
- Atman, C., Cardella, M., J. Turns, & Adams, R. S. (2005). Comparing Freshman and Senior Engineering Design Processes. *Design Studies*, 26(4), 325–357.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), Article 2. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bartholomew, S., Mentzer, N., Jones, M., Sherman, D., & Baniya, S. (2022). Learning by evaluating (LbE) through adaptive comparative judgment. *International Journal of Technology and Design Education*, 32(2), 1191–1205. <https://doi.org/10.1007/s10798-020-09639-1>
- Bartholomew, S., Strimel, G., Garcia Bravo, E., Zhang, L., & Yoshikawa, E.

- (2018). Formative feedback for improved student performance through adaptive comparative judgment. 2018 ASEE Annual Conference & Exposition, Salt Lake City, Utah. <https://doi.org/10.18260/1-2--30531>
- Bartholomew, S., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, 29(2), 363–385. <https://doi.org/10.1007/s10798-018-9442-7>
- Bartholomew, S., Yauney, J., Mentzer, N., & Thorne, S. (2024). Investigating the impacts of differentiated stimulus materials in a learning by evaluating activity. *International Journal of Technology and Design Education*, 34(4), 1561–1577. <https://doi.org/10.1007/s10798-023-09871-5>
- Bartholomew, S., Zhang, L., Garcia Bravo, E., & Strimel, G. J. (2019). A tool for formative assessment and learning in a graphics design course: Adaptive comparative judgement. *The Design Journal*, 22(1), 73–95. <https://doi.org/10.1080/14606925.2018.1560876>
- Buckley, J., Canty, D., & Seery, N. (2022). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies*, 41(2), 313–331. <https://doi.org/10.1080/03323315.2020.1814838>
- Canty, D. (2012). *The impact of holistic assessment using adaptive comparative judgment of student learning*. University of Limerick.
- Cash, P., & Kreye, M. (2018). Exploring uncertainty perception as a driver of design activity. *Design Studies*, 54. <https://doi.org/10.1016/j.destud.2017.10.004>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Academic Press.
- Collins, A. (n.d.). *Cognitive Apprenticeship* [The International Society of the Learning Sciences]. <https://www.isls.org/research-topics/cognitive-apprenticeship/>
- Collins, A., Brown, J., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, 6, 38–46.
- Crismond, D. P., & Adams, R. S. (2012a). The Informed Design Teaching and Learning Matrix. *Journal of Engineering Education*, 101(4), 738–797. <https://doi.org/10.1002/j.2168-9830.2012.tb01127.x>
- Crismond, D. P., & Adams, R. S. (2012b). The Informed Design Teaching and Learning Matrix. *Journal of Engineering Education*, 101(4), 738–797. <https://doi.org/10.1002/j.2168-9830.2012.tb01127.x>
- Cross, N. (2006). *Designerly Ways of Knowing*. Springer-Verlag. <https://doi.org/10.1007/1-84628-301-9>
- Dennen, V. P., & Burner, K. J. (2008). The cognitive apprenticeship model in educational practice. In *Handbook of research on educational communications and technology* (pp. 425–439). Routledge.

- Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., & Leifer, L. J. (2005). Engineering Design Thinking, Teaching, and Learning. *Journal of Engineering Education*, 94(1), 103–120. <https://doi.org/10.1002/j.2168-9830.2005.tb00832.x>
- Fila, N., Purzer, S., & Mathis, P. (2014). *I'm not the creative type: Barriers to student creativity within engineering innovation projects*. ASEE Annual Conference and Exposition, Indianapolis, IN. <https://peer.asee.org/i-m-not-the-creative-type-barriers-to-student-creativity-within-engineering-innovation-projects.pdf>
- Gensemer, A., & Caron, D. (2010). *Foundations of Technology* (3rd ed.). International Technology and Engineering Educators Association.
- Hawe, E., Dixon, H., & Hamilton, R. (2021). Why and how educators use exemplars. *Journal of University Teaching and Learning Practice*, 18(3). <https://doi.org/10.53761/1.18.3.10>
- Hill, B. (2006). *Introduction to engineering design notes*. University of Maryland Baltimore County. <https://umbc.edu/pltw/resources/Inventor3.pdf>
- International Technology and Engineering Educators Association (ITEEA). (2020). Standards for technological and engineering literacy: The role of technology and engineering in STEM education.
- Jackson, A., Bartholomew, S., & Mentzer, N. (2022). Priming the design process: Activating and characterizing students' critical thinking in design. Mississippi Valley Technology Teacher Education Conference. <https://par.nsf.gov/biblio/10433482-priming-design-process-activating-characterizing-students-critical-thinking-design>
- Jackson, A., & Strimel, G. (2018). Toward a matrix of situated design cognition. CTETE Research Monograph Series, 1(Summer), 49–65. <https://doi.org/10.21061/ctete-rms.v1.c.3>
- Jang, H. (2016). Identifying 21st Century STEM Competencies Using Workplace Data. *Journal of Science Education and Technology*, 25(2), 284–301.
- Kelley, T. (2011). Engineer's notebook – A design assessment tool. *Technology and Engineering Teacher*, 70(7), 30–35.
- Kelley, T., & Knowles, J. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, 3(1). <https://doi.org/10.1186/s40594-016-0046-z>
- Lee, W., Mentzer, N., Jackson, A., Bartholomew, S., & Clevenger, A. (2024). Defining and Evaluating Argumentation Quality in the Context of Design Thinking: Using High School Students' Design Critiques from Foundational Engineering Courses. *Design and Technology Education: An International Journal*, 29(3), 11–35. <https://doi.org/10.24377/DTEIJ.ARTICLE2530>
- Lee, W., Mentzer, N., Jackson, A., Bartholomew, S., & Thorne, S. (2025). A thematic analysis of high school students' scientific argumentation of what

- constitutes a ‘better’ engineering design journal. *International Journal of Technology and Design Education*. <https://doi.org/10.1007/s10798-025-09984-z> Lomask, M., Crismond, D., & Hacker, M. (2018). Using Teaching Portfolios to Revise Curriculum and Explore Instructional Practices of Technology and Engineering Education Teachers. *Journal of Technology Education*, 29(2), Article 2.
- Lucas, B., & Hanson, J. (2016). Thinking Like an Engineer: Using Engineering Habits of Mind and Signature Pedagogies to Redesign Engineering Education. *International Journal of Engineering Pedagogy (iJEP)*, 6(2), 4. <https://doi.org/10.3991/ijep.v6i2.5366>
- Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education*, 95(5), 877–907. <https://doi.org/10.1002/sce.20441>
- Mentzer, N. (2014). High school student information access and engineering design performance. *Journal of Pre-College Engineering Education Research*, 4(1), Article 1. <https://doi.org/10.7771/2157-9288.1074>
- Mentzer, N., Becker, K., & Sutton, M. (2015). Engineering Design Thinking: High School Students’ Performance and Knowledge. *Journal of Engineering Education*, 104(4), 417–432. <https://doi.org/10.1002/jee.20105>
- Mentzer, N., Lee, W., Jackson, A., & Bartholomew, S. (2024). Learning by Evaluating (LbE): Promoting meaningful reasoning in the context of engineering design thinking using Adaptive Comparative Judgment (ACJ). *International Journal of Technology and Design Education*, 34(3), 1145–1169. <https://doi.org/10.1007/s10798-023-09853-7>
- Moore, T. J., Tank, K. M., Glancy, A. W., & Kersten, J. A. (2015). NGSS and the landscape of engineering in K-12 state science standards. *Journal of Research in Science Teaching*, 52(3), 296–318.
- National Academy of Engineering, & National Research Council. (2009). *Engineering in K-12 Education: Understanding the Status and Improving the Prospects* (p. 12635). National Academies Press. <https://doi.org/10.17226/12635>
- National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. National Academies Press.
- National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. National Academies Press.
- National Science Board. (2023). *Explorations in STEM PreK-12 Education Final Report*. National Science Foundation. https://www.nsf.gov/nsb/publications/2023/ESKE_final_report.pdf
- Next Gen Science, (n.d.). *Transforming Science Assessment: Challenges and Recommendations for States*. Retrieved October 21, 2025, from

- <https://www.nextgenscience.org/transforming-science-assessment-challenges-and-recommendations-states>
- NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.
- Osmond, J., & Tovey, M. (2015). The Threshold of Uncertainty in Teaching Design. *Design and Technology Education*, 20(2), 50–57.
- Paretti, M. C. (2008). Teaching Communication in Capstone Design: The Role of the Instructor in Situated Learning. *Journal of Engineering Education*, 97(4), 491–503. <https://doi.org/10.1002/j.2168-9830.2008.tb00995.x>
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), Article 3.
- RM plc. (2025). *RM Compare* [Computer software] Validated human consensus at scale. <https://compare.rm.com/>
- Sadler, R. (2002). Ah! ... So That's "Quality". In P. Schwartz & G. Webb (Eds.), *Assessment: Case studies, experience and practice from higher education* (pp. 130–136). RoutledgeFalmer.
- Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education*, 29(4), 701–715. <https://doi.org/10.1007/s10798-018-9468-x>
- Seery, N., & Canty, D. (2017). *Assessment and learning: The proximal and distal effects of comparative judgment*. In M. J. de Vries (Ed.), *Handbook of technology education* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-38889-2_54-1.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *Journal of the Learning Sciences*, 3(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

- Thorne, S., Mentzer, N., Strimel, G., Bartholomew, S., & Ware, J. (2024). Learning by Evaluating: An Exploration of Optimizing Design-Based Instruction. *Journal of Technology Education*, 35(2), 53–80. <https://doi.org/10.21061/jte.v35i2.a.4>
- To, J., Panadero, E., & Carless, D. (2022). A systematic review of the educational uses and effects of exemplars. *Assessment & Evaluation in Higher Education*, 47(8), 1167–1182. <https://doi.org/10.1080/02602938.2021.2011134>
- Tracey, M. W., & Hutchinson, A. (2016). Uncertainty, reflection, and designer identity development. *Design Studies*, 42, 86–109. <https://doi.org/10.1016/j.destud.2015.10.004>
- Watson, A. (2025). Employment and wages for alternate definitions of science, technology, engineering, and mathematics (STEM) occupations. *Beyond the Numbers: Employment & Unemployment*, 14(6). <https://www.bls.gov/opub/btn/volume-14/stem-alternate-definitions.htm>
- Wu, H.-C., Shen, P.-D., Chen, Y.-F., & Tsai, C.-W. (2016). Effects of Web-based Cognitive Apprenticeship and Time Management on the Development of Computing Skills in Cloud Classroom: A Quasi-Experimental Approach. *International Journal of Information and Communication Technology Education*, 12(3), 1–12. <https://doi.org/10.4018/ijicte.2016070101>

About the Authors

Nathan Mentzer (nmentzer@purdue.edu) is Professor at Purdue University, West Lafayette, Indiana USA in the Purdue Polytechnic College with a joint appointment to the College of Education. <https://orcid.org/0000-0001-9788-7901>

Abdul Moeed Asad (asadab@iu.edu) is a PhD student at Indiana University Bloomington, Indiana USA, studying design theory and pedagogy. <https://orcid.org/0009-0000-3580-8788>

Wonki Lee (lee01key@gmail.com) is the Associate Director of Educational Assessment Services & Initiatives at the University of California, Los Angeles, California USA. <https://orcid.org/0000-0002-0007-1023>

Andrew Jackson (andrewjackson@uga.edu) is an Associate Professor of Workforce Education at the University of Georgia, Athens, GA USA. <https://orcid.org/0000-0003-2882-3052>

Scott Bartholomew (scottbartholomew@byu.edu) is an Associate Professor of Technology and Engineering Education at Brigham Young University in Provo, Utah, USA. <https://orcid.org/0000-0002-1680-2433>