International Journal of Technology and Design Education (2024) 34:1145–1169
https://doi.org/10.1007/s10798-023-09853-7

Learning by Evaluating (LbE): promoting meaningful reasoning in the context of engineering design thinking using Adaptive Comparative Judgment (ACJ)

Nathan Mentzer[1] · Wonki Lee[2] · Andrew Jackson[3] · Scott Bartholomew[4]

Accepted: 7 September 2023 / Published online: 11 October 2023

Abstract
Adaptive comparative judgment (ACJ) has been widely used to evaluate classroom artifacts with reliability and validity. In the ACJ experience we examined, students were provided a pair of images related to backpack design. For each pair, students were required to select which image could help them ideate better. Then, they were prompted to provide a justification for their decision. Data were from 15 high school students taking engineering design courses. The current study investigated how students' reasoning differed based on selection. Researchers analyzed the comments in two ways: (1) computer-aided quantitative content analysis and (2) qualitative content analysis. In the first analysis, we performed sentiment analysis and word frequency analysis using natural language processing. Based on the findings, we explored how the design thinking process was embedded in student reasoning, and if the reasoning varied depending on the claim. Results from sentiment analysis showed that students tend to reveal more strong positive sentiment with short comments when providing reasoning for the selected design. In contrast, when providing reasoning for those items not chosen, results showed a weaker negative sentiment with more detailed reasons. Findings from word frequency analysis showed that students valued the function of design as well as the user perspective, specifically, convenience. Additionally, students took aesthetic features of each design into consideration when identifying the best of the two pairs. Within the engineering design thinking context, we found students empathize by identifying themselves as users, define user's needs, and ideate products from provided examples.

Keywords Adaptive Comparative Judgment (ACJ) · Learning by Evaluating (LbE) · Claim, Evidence, Reasoning (CER) model · Design thinking · Technology education

✉ Nathan Mentzer
nmentzer@purdue.edu

[1] Purdue Polytechnic Institute and College of Education, West Lafayette, IN, USA

[2] Center for Instructional Excellence, Purdue University, West Lafayette, IN, USA

[3] Department of Workforce Education and Instructional Technology, University of Georgia, Athens, GA, USA

[4] Brigham Young University, Technology and Engineering Studies, Provo, UT, USA

Introduction

Recent studies in STEM education highlighted an emphasis on peer critique and feedback, as both formative assessment and learning strategy. Adaptive comparative judgment (ACJ; Pollitt & Whitehouse, 2012; Pollitt, 2012a, 2012b) was developed based on the law of comparative judgment (Thurstone, 1927), which allows judges to compare items which are then ranked based on the results. ACJ is a modification on traditional Comparative Judgment (CJ) which exploits the power of adaptivity to expedite the comparison process (Pollitt, 2012b). Other work has also demonstrated that ACJ can be used as a formative assessment, which challenges the evaluator to discern qualitative differences between artifacts (Pollitt, 2012b) during the learning process. In the context of ACJ, judges are provided pairs of work, then, they make decisions about which one is better. From multiple comparisons, a measurement scale provides relative quality of the artifacts. Additionally, ACJ also can be incorporated in educational contexts as a learning mechanism. Along with the written justification for the decisions, studies found that ACJ can be implemented as a meaningful assessment and feedback tool, in which students can improve learning and achievement (Bartholomew & Jones, 2022; Bartholomew et al., 2019). In this case, our research team has named the priming process *Learning by Evaluating* (LbE) which stimulates and promotes meaningful learning for students (Bartholomew, 2021; Bartholomew & Yauney, 2022; Bartholomew et al., 2020). Beginning student efforts through LbE builds foundational understanding and has been shown to produce meaningful learning for students (Bartholomew, 2021; Bartholomew & Yauney, 2022; Bartholomew et al., 2020).

For learning purposes, LbE has shown promise in facilitating student learning and growth (Bartholomew et al., 2018a, 2018b, 2019; Canty et al., 2017; Sluijsmans et al., 1998a; Strimel et al., 2021). Additional work has shown that students' decision-making during the peer assessment process promotes critical thinking (Jackson et al., 2022; Sluijsmans et al., 1998a, 1998b) and an analytical approach to design (Nicol et al., 2014). In multiple acts of making judgment, students think about not only the provided example but also their own work. Thus, LbE is beneficial in that it promotes active engagement in critical thinking, applying criteria, reflecting, and learning transfer derived from this process (Nicol et al., 2014). Especially, regular opportunities for students to participate in making feedback significantly enhance learning outcomes (Sadler, 2010). Additionally, Cowan (2010) asserted the importance of systemic, objective, and sound criteria-based decision and reasoning for making effective evaluative judgment. Therefore, the learning benefits of LbE may be increased as students engage in providing rationales for their claims.[1]

These findings around LbE are well aligned with the *Claim, Evidence and Reasoning* (CER) model developed by McNeill et al., (2015) who position CER as a scientific argument model. A *claim* provides an answer to the question or problem. *Evidence* is data that supports the claim (e.g., measurements, observation, predetermined criteria), and *reasoning* explains why and how the evidence supports the claim using principles. In the LbE context, students engage in making a claim when they are tasked with selecting a better example from a pair of options. Student explanations are a rationale for their decision

[1] In the present study, students did not write feedback for the purpose of providing it to their peers. Instead, they evaluated example artifacts where reasoning provided in the comment functioned as a primer task for students. Hence, we will use the term 'comment' to designate the output of reasoning. Thus, eliciting feedback will concentrate on activating students as owners of their own learning, helping students articulating what success looks like, and critically reflecting on their own learning.

which may include evidence and their reasoning with criteria. The current study investigated the extent to which this reasoning differed by selecting and not selecting one of the pair of items displayed in each judgment, and how the reasoning differed in terms of word sentiment, usage, word frequency. Further, we compared the contents of the reasoning qualitatively in terms of design thinking. More specifically, we explored how students' reasoning incorporated the learned principles of the design thinking process and if there was any difference based on their claim (i.e., reasoning for the selected option vs. not selected option).

Previous studies

Three major strands of studies are introduced in this section. First, the origin and the notion of ACJ will be introduced. Then, the incorporation of ACJ and its learning benefits will be presented. Under this strand, LbE—which designates a specific application of ACJ and learning benefits coming from taking part in ACJ—is also discussed. Finally, we elaborate on the CER model and its implication in the STEM education context.

Adaptive Comparative Judgment (ACJ)

Adaptive comparative judgment was originally designed as an approach to assessment and evaluation. It was established on the "Law of Comparative Judgment" (Thurstone, 1927) by Louis Thurstone who sought to measure psychological values with discriminable differences. He presented a series of articles raising a concern with the methods of measurements of psychological perceptions such as mind, social attitudes, intelligence, and values (Thurstone, 1927, 1931, 1941, 1954). Through this work, he yielded a meaningful finding of the "Law of Comparative Judgment" which can be used to measure the physical intensities, as well as quality of psychological values, through simultaneous and successive contrasts (Thurstone, 1927).

Later, (Pollitt, 2012b) outlined the possibility of adaptive comparative judgment (ACJ) which is applicable to various educational assessments and could provide high reliability and validity compared to the conventional assessment (i.e., operational marking) by graders (Mentzer et al., 2021; Pollitt, 2012b). Pollitt incorporated an adaptive attribute to the comparative judgment by pairing similarly-ranked items in the process of judging (as opposed to random comparisons) to maximize the information gained from each comparison. Also, ACJ leverages a holistic statement which guides the judges to make a professional judgment (Pollitt, 2012b) as to which item—of the pair displayed—demonstrates the highest quality. Judges make a comparison based on the provided holistic statement with overarching reasoning (e.g., "Which artifact demonstrates a better ideation process?").

Learning by evaluating (LbE) in STEM education

Several recent studies have incorporated adaptive comparative judgment (ACJ) into STEM education setting, especially in project-based design thinking processes (Bartholomew et al., 2018a, 2018b, 2019, 2020; Dewit et al., 2021; Strimel et al., 2021). Bartholomew et al., (2018a, 2018b) examined ACJ to evaluate the middle school students' learning through an open-ended problem assigned in a technology and engineering education course. They found ACJ to be an effective assessment with reliability, validity, feasibility, and practicality. Results from assessment using ACJ also correspond to traditional marking results in project-based learning (Bartholomew et al., 2018a, 2018b) but this innovation presents other opportunities. Not only was ACJ an effective measure, but, when students participate in ACJ, they can experience learning without necessarily increasing teacher workload (Nicol et al., 2014).

Engaging students in the ACJ process has numerous intellectual merits (Collins et al., 2018). Critiquing, evaluating, and providing reasoning can foster critical thinking (Hokanson, 2012; McNeill & Krajcik, 2008). Other studies have asserted that student reasoning in the process of peer feedback has benefits such as a better understanding of learning goals (Stefani, 1994), more advanced understanding of assignments (Nicol et al., 2014), and higher achievements (Gielen & De Wever, 2015). Furthermore, the ACJ process gives an intuitive method to involve students in the assessment process as judges. Students reported ACJ as a fun and enjoyable task and students are keen to discuss their judgments (Jones & Wheadon, 2015). Later, Bartholomew et al. (2019) implemented peer-evaluated ACJ in an open-ended design setting. In this study, students participated as judges to make assessments for their peer's design projects. A few studies delve into student-based ACJ and its educational benefits. Strimel et al. (2021) explored students' ACJ process on their peer work and results showed that students reflected on their own design during the ACJ and thus gain future ideas to implement into their designs. Dewit et al. (2021) incorporated ACJ in the higher educational context, a product-service-system design project for first year master's students, and found that these students could incorporate analytical thinking and metacognition, which improved their learning experiences. Collectively, the findings illuminated possibilities of using ACJ as an effective means for both teaching and learning.

Extending findings from Bartholomew et al. (2019), Bartholomew et al. (2020) introduced a new term to describe the use of ACJ as a primer for learning: *Learning by Evaluating* (LbE). LbE engages students as judges using the ACJ platform, but facilitates design thinking insights at the start of a project, rather than waiting until students have progressed significantly further with their work. Moreover, instead of evaluating peer work, the examples used in LbE are selected by the instructor to target specific learning objectives or project elements. Working in this paradigm, we hypothesize that students will be oriented to the project and that the comparisons made will further their critical thinking and decision making when approaching open-ended problems (Jackson et al., 2022).

CER (Claim-Evidence-Reasoning) framework

The process of making and justifying decisions, also called argumentation, is central in engineering (Wilson-Lopez et al., 2020) and design (Daly et al., 2012). The claims, evidence, and reasoning (CER) framework, which emerged from research and instruction in science education, describes three central components of scientific thinking and argumentation that are applicable in these knowledge-construction settings (McNeill & Krajcik, 2008). This framework has been used in instruction, to support students' decision-making and explanation processes. By using the CER framework to craft an argument, students gain a deeper understanding of the learning content, because their explanations focus on bringing in their scientific background knowledge (Fielding-Wells, 2016). Even in complex inquiry-based problems, Gillies (2020) supported this claim by asserting that students can use reasoning based on disciplinary knowledge. Context also plays a central role in CER framework, in defining the question, design, or problem within which the claims are situated (NGSS Standards, 2013).

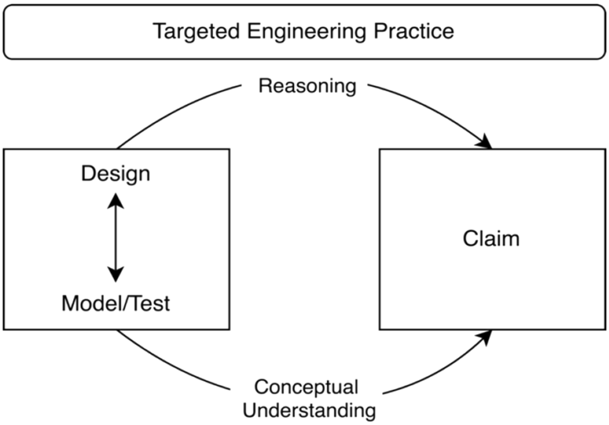

Fig. 1 Ways of thinking in engineering: using reasoning to make claims (Slavit et al., 2021)

Based on the CER model (Slavit et al., 2021), we have chosen to concentrate on the role of claim-making and reasoning that occurs in engineering design thinking activities. In the design thinking context, students learn while designing, and can begin to make and explain knowledge-driven decisions (Crismond & Adams, 2012). They learn by doing (e.g., iteration, drawing, and troubleshooting), and use this knowledge from experiments to make a better, evolving design decision (Perkins et al., 1995). Yet, in engineering design thinking argumentation (see Fig. 1), the claims are tentative because the goal is to create and enact the most appropriate model for the contextualized problem (Slavit et al., 2021). Viability based on experiment and user-response are two major criteria that drive the claim-making process. Still, each claim, the statement or a conclusion to a problem, should be supported by justification based on data and explanation as in the CER framework and scientific argumentation (McNeill & Krajcik, 2008). Reasoning logically connects between claim and evidence, by expounding how those relationships can be established (McNeill & Krajcik, 2008; Toulmin, 2003; Van Eemeren et al., 2019).

In the current study, we determined to delve into students' reasoning in engineering design settings, when incorporating LbE. The learning objectives and artifacts used for the experiences were predetermined and meant to support students' design work, particularly in terms of enhancing their ideation. As seen in Fig. 2, by selecting option A or option B, they make an implicit claim that "Option A is better" or "Option B is better." Then, they are further prompted to answer a question: "Why you have chosen A (or B)?" While providing comments about their judgment decision, students need to provide reasoning for their claim. These claims and explanations provide the basis for our investigation which centered on student reasoning and content.

Methods

Research questions

The current study aimed to examine students' reasoning from LbE in the context of design thinking. Specifically, our focus was on the reasoning provided in the comments section of LbE done by students working on a backpack design challenge. Considering that students provide reasoning for either the 'selected' option (e.g., "I like option A because it looks

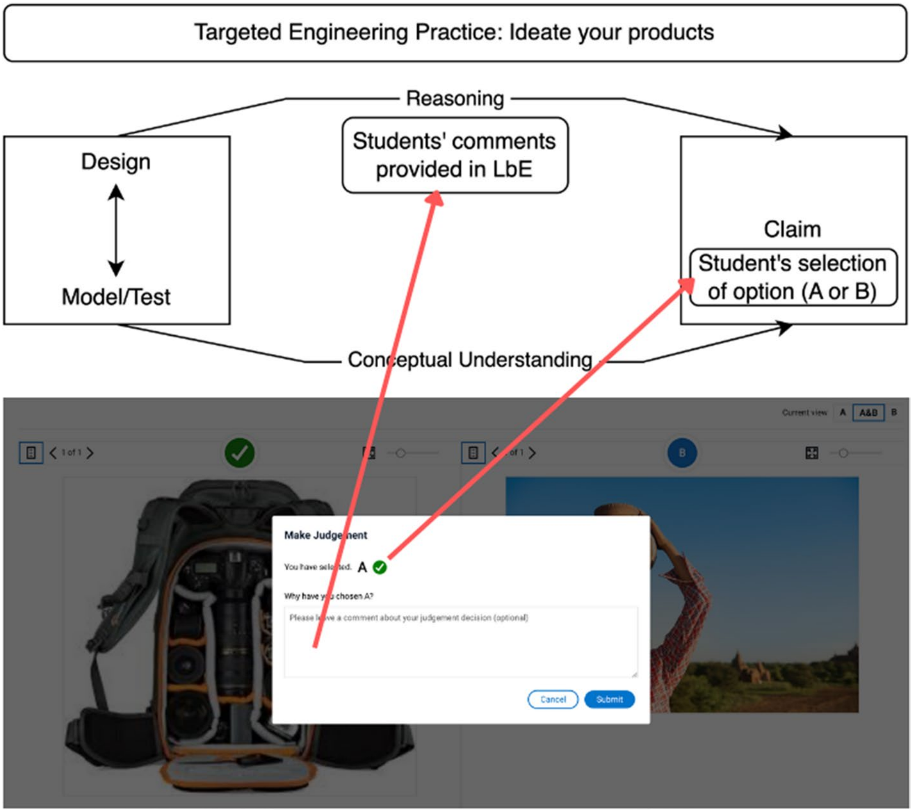
Targeted Engineering Practice: Ideate your products
Reasoning
Students' comments provided in LbE
Design
Model/Test
Claim
Student's selection of option (A or B)
Conceptual Understanding
Current view A A&B B
1 of 1
B
Make Judgement
You have selected: A
Why have you chosen A?
Please leave a comment about your judgement decision (optional)
Cancel
Submit

Fig. 2 CER Model implication in the LbE process

fancier.”) or ‘not selected’ option (e.g., choosing option A because “Option B cannot carry a thing."), our intent was to determine if the content of the reasoning differed, and how it differed. In this empirical analysis, the following research questions were used to compare these different reasonings based on selection (i.e., selected and not selected).

RQ 1. Is there a statistically meaningful difference between the scores of sentiment analysis (e.g., positive, negative, neutral, and compound scores) based on reasoning target?

RQ 2. How are the contents of reasoning situated in the design thinking context? In other words, what are the design thinking components students incorporated in their reasoning?

Research context

This study, part of a larger project, was embedded in a 9th grade Engineering design course with five units during the school year. Students learned design experiences in the early units of lessons, and then worked to apply what they learned later in the unit. Researchers partnered with a large, urban district near Atlanta, Georgia, United States. The district serves a total of 93,703 students from a very diverse population (DeKalb County School

District, 2021). Students and their families speak over 185+languages and represent over 155+nations. The majority of the students are underrepresented minorities (59% Black, 20% Hispanic), and 48.2% of students are eligible for free and reduced lunch (DeKalb County School District, 2021).

Teachers in this study received professional development training multiple times as part of their *Engineering by Design* certification (International Technology and Engineering Educators Association, 2018) for the course, as well as training in LbE implementation in the classroom to ensure fidelity of use in our larger project. Teachers then introduced the design projects, led classroom discussions among the students, facilitated our LbE research, promoted skill development in the design process, and tracked student progression. With the guidance of teachers, students learned basic concepts of a design thinking process and its components. Further, they were motivated to participate in hands-on tasks through project-based learning of design thinking, and could address meaningful, real-world problems fostering critical thinking.

Participants

Prior to the data collection, consent and assent forms were distributed to students, teachers, and parents/guardians, as required by the researchers' University Institutional Review Board and the school district's research coordinators. This study focused where teachers were delivering a similar challenge to their introductory Engineering design classes—to complete a backpack design. All 140 students enrolled in the course across three schools were invited to participate in the study. Out of three schools, two teachers consented (66.67% return rate). Out of 140 students, 32 students completed assent forms (22.86% return rate) and 37 parents consented (26.43%). Thus, only 15 students were fully consented by school, teacher, and parents. Students were in their 9th grade engaged in the Engineering design course during Fall 2021 and Spring 2022.

Task: Backpack comparison

The goal of the current task was to promote the students' ideation process as part of design thinking. While identifying which one of the pair to choose, students were required to select which example helped or inspired them to ideate better. Our holistic statement for the current task was:

What makes a good backpack? How could each item inspire you to solve the problem. Which item best assists you with your design and why?

As an example of a comparison, in Fig. 3, students can think about which option is better and why it is better for their ideation of a new product design.

Researchers uploaded a total of 39 different design artifacts from different sources related to the ideation of effective backpack design. To help in student brainstorming, images included various items such as actual backpacks, or other images related to backpacks that might carry things or hold things. During the activity, teachers introduced LbE with three steps: a student orientation, the LbE comparisons, and a post-session debrief. For instance, teachers provided a detailed and actionable problem statement to address the current problem of designing 'better' and 'inspiring' backpacks to users. Students then viewed pairs of images. In each case, they were required to select which they believed to be the better option (claim) to help them consider ways to improve their own backpack designs. Right after their decision, students were required to justify their choice of one item

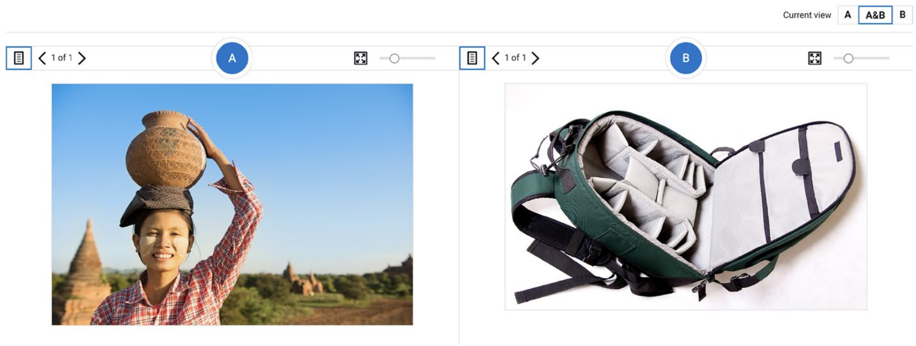

Fig. 3 Screenshot of Backpack comparison session. "Left (**A**): ©WONG SZE FEI/ Adobe Stock #99,606,465", "Right (**B**): ©cegli/ Adobe Stock #27,278,935"

over another. Students either could provide their reasoning for 'selecting' the better image of the pair or describe why they didn't select the other option. After the completion of the judgments, teachers transitioned to a group discussion about which design artifact was better at inspiring ideation and why.

Each student made an average of 10 comparisons with a minimum of 3 and a maximum of 20 judgments. After removing responses which do not fit the analysis criteria (e.g., not providing comments, writing something unrelated, including abusive/inappropriate language), a total of 145 comments from the judgments were collected.

Research design

Content analysis

According to Neuendorf (2017), content analysis is a systematic and objective analysis of the characteristics of a message in context which can include both "human-coded analysis and computer-aided text analysis" (p. 19). Content analysis helps interpret meaning from the content of text data, and allows researchers to infer about the "states or properties of the sources of the analyzed data" (Krippendorff, 2018, p. 20). It postulates concepts, attitudes, beliefs, intentions, emotions, and cognitive processes manifested in text; thus, it sets those attributes as a natural target for successful content analysis (Krippendorff, 2018). Both qualitative and quantitative approaches may be used in content analysis. In quantitative analysis, text is decontextualized to show patterns with reliability and validity. As compared to quantitative inquiry, qualitative inquiry aims to understand a phenomenon, rather than generalize the phenomenon (Forman & Damschroder, 2007), including thematic description.

Since the first implementation in the field of content analysis, computer-aided content analysis (Sebeok & Zeps, 1958) is gaining its popularity, especially in the current big-data and machine-learning environment. It has proved high reliability, validity, and efficiency in different research settings (Su et al., 2017). However, computer-aided content analysis has key limitations, such as not always tapping the true meaning behind the relationships between the words (Krippendorff, 2018; Matthes & Kohring, 2008; Su et al., 2017). In response to the limitations noted above, we adopted a hybrid approach by combining computation and human-based methods to take advantage of researchers' insights in

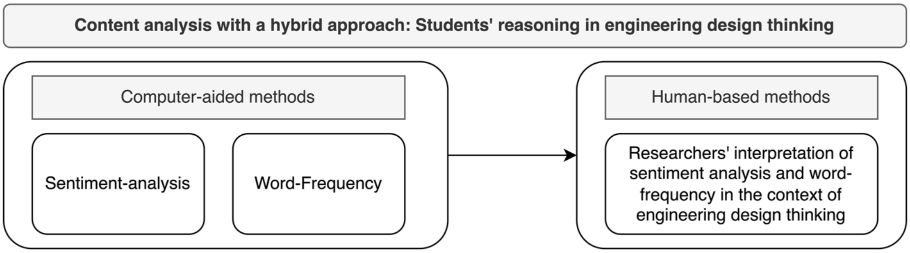
Content analysis with a hybrid approach: Students' reasoning in engineering design thinking
Computer-aided methods
Sentiment-analysis
Word-Frequency
Human-based methods
Researchers' interpretation of sentiment analysis and word-frequency in the context of engineering design thinking

Fig. 4 Content analysis with a hybrid approach

the analysis of text datasets (Su et al., 2017). As seen in Fig. 4, based on the text-mining approach, researchers utilized an additional interpretation to recognize underlying meanings embedded in the result of computer-aided content analysis.

Researchers agreed upon utilizing students' reasoning in the comment section of each judgment to analyze their reasoning when making a claim in an engineering design thinking context. To explore differences of reasoning between 'selected' and 'not selected' options, we performed sentiment analysis and word frequency analysis. First, we labeled the data based on the justification of students. Then, we performed sentiment analysis to see if the perceived sentiment is different for the selected or the not selected option. A *t* test was used to identify if there was a statistically significant difference on sentiment between reasoning targets. Followed by this, we performed content analysis with word frequency. It revealed how word count was different between selected and not selected decision reasonings. Finally, researchers explored the students' thought process by identifying specific characteristics and then comparing those responses to an engineering design mindset. For the analysis, we used Python 3.9.7 within Visual Studio Code (version 1.71). Libraries used will be stated in each subsection.

Data labeling

Because student claims and justifications were made holistically, not on either item in the comparison, it was necessary to review and label whether the justification was based on the selected or not selected option. After the initial labeling from one of the research team members who led this analysis, each student justification was manually reviewed and coded by the research team based on our interpretation of the target of the comment. As seen in the examples in Table 1, when students provided reasoning about the selected option, we coded them as a 0. In contrast, when it was about not selected option, we coded them into 1.

Three of the comments included reasonings related to both options in the comparison. Previous studies found that including mixed reviews in classification creates a problem when scoring words (Akhtyamova et al., 2018; Dave et al., 2003; Dey & Haque, 2008, 2009). We decided to separate these comments out due to the small number of comments with mixed reasoning (N = 3) or about the not selected option (N = 16). Hence, we separated mixed comments (see Table 2) into two comments and coded them in the same way as before, based on whether the comment referred to the item selected or not selected (see Table 3). Therefore, the total number of comments slightly increased (N = 148).

Table 1 Example codes for reasoning target, based on claim

Comments	Selected*	Code
Helps me show that straps could be involved in my design	Y	0
No one would buy the other one	N	1
This looks very impractical, but somewhat stylish	Y	0
The other one is hideous, uncomfortable, and heavy	N	1
It shows that the strap is adjustable so you can change to your likings	Y	0

Selected* means if it is a reasoning about the selected option

Overall, students provided more reasoning for the selected option (*N* = 129) than the not selected options (*N* = 19). Our next step involved application of computer-aided content analysis via natural language processing.

Data preprocessing using natural language processing

After finishing the labeling, researchers processed the natural language into the vectorized text. Natural language processing incorporates computational techniques to learn, understand, and process human language (Hirschberg & Manning, 2015; Nadkarni et al., 2011). We removed punctuation, stemmed, lemmatized, and converted the text to the features into the model below (see Table 4). Specifically, we used lemmatization and count vectorizer function from NLTK packages (Bird et al., 2009). Using lemmatization, we processed words into their normalized form, for example to simplify 'straps' to 'strap' (Plisson et al., 2004). Then, we removed stop words—common words like 'I', 'me', 'was'—to find the context or the true meaning of decision criteria. Finally, count vectorizer was incorporated to quantify the words, by taking the frequency of words into account for the prediction. After initial preprocessing, researchers decided to add six more stop words to remove words repeated without significant meaning in the current research context (i.e., 'option', 'b').[2] After preprocessing, reasonings about the selected option had a lower character count (*M* = 54.40, *SD* = 33.25) and word count (*M* = 10.22, *SD* = 6.30) compared to the character count and word count for reasonings about the not selected option (*M* = 71.37, *SD* = 57.85; *M* = 13.84, *SD* = 10.66, respectively).

Sentiment analysis

Sentiment analysis was initiated to gauge opinions and identify sentiments (Mejova, 2009; Prabowo & Thelwall, 2009; Zhang et al., 2018). It often provides a polarity, dichotomized into two: positive and negative (Mejova, 2009). In this study, we used NLTK's Vader analysis, which classifies the texts into different sentiments as a range (Bird et al., 2009). Sentiment intensity analyzer was implemented, which yields scores of positive, negative, neutral, and compound results. The compound score is the normalized sum of all positive, negative, and neutral scores. Most extreme negative text (e.g., 'I hate A.') is scored as -1,

[2] 'A' in the 'Option A' was not included in the new stop words because 'a/A' is already included in the stop words due to the article 'a'.

Table 2 Example codes for comments when participants provided mixed reasoning

Comments	Selected*	Code
Backpacks are way more comfortable and better in ALL ways than a purse; A backpack will lay on your shoulders with straps while a purse will irritate the skin on the inside of your elbow. It's just true	N/A	N/A

Selected* means if it is a reasoning about the selected option

and most extreme positive text (e.g., 'it contains all of the things you need for school work food etc.') is scored as + 1. Thus, the closer the compound score is to + 1, the more the reasoning is considered positive; conversely, the closer the compound score is to –1, the more the reasoning is considered negative. In Table 5, #Ex1—#Ex3 shows how the sentiment score is assigned to students' reasoning.

Following sentiment analysis, we performed *t* tests to compare the difference of sentiments between two reasoning targets. Each sentiment score—positive, neutral, negative, and the compound result that represents an overall sentiment for the reasoning—was compared by group for the reasoning target: selected or not selected.

Word Frequency

Word frequency was also incorporated as part of the content analysis. Word frequency in a document provides insights about the patterns of word used and the content of a text (Baayen, 2001). After studying word frequency, we decided to present the top 10 most frequently used words.

Results

In this section, findings from computer-aided content analysis (i.e., sentiment analysis, word frequency) are presented. Then, researchers introduce conclusions from human-based content analysis, which elaborated on the findings from computer-aided content analysis.

Students' perceived sentiment in reasoning based on claim

As seen in Table 6, students used words with more positivity ($M=0.26$, $SD=0.29$) when providing reasoning for the selected option, while they used more negative words ($M=0.10$, $SD=0.28$) when providing their reasoning for not selecting an option. In terms of neutral sentiment score, we found reasoning for selected options showed lower neutral sentiment scores ($M=0.73$, $SD=0.29$) compared to reasoning related to the not selected option ($M=0.78$, $SD=0.30$). The differences in each reasoning sentiment (in isolation) also explains the higher compound score of the reasoning about the selected group ($M=0.22$, $SD=0.25$) in comparison to the reasoning about the not selected group ($M=0.00$, $SD=0.28$). These sentiment scores showed that students had a tendency to express strong positivity for the selected options, while providing less strong negativity toward the option which was not selected (see Fig. 5).

Table 3 Example mixed coded comments after separating them into two reasonings

Comments	Selected*	Code
Backpacks are way more comfortable and better in ALL ways than a purse; A backpack will lay on your shoulders with straps	Y	0
A purse will irritate the skin on the inside of your elbow. It's just true	N	1

Selected* means if it is a reasoning about the selected option

The results from the *t* test indicated that differences between the scores were statistically significant for the positive sentiment score, $t(146) = 5.87$, $p < 0.01$. There was no statistically significant difference found in negative sentiment score and neutral sentiment score using the threshold level of $p < 0.01$. In other words, whether student comments related to the selected option or the not selected option, students used similar levels of neutral and negative words in their reasoning. However, students tended to use strong positive words when providing reasoning for selected options.

In Table 7, five comments with the highest negative sentiment score are presented (Neg1–Neg5). Because not all the words in the comment included negative sentiment, we could not find 1.0 point negative statement from the students' reasoning. Even though students provided reasoning for the selected option, the negative sentiment score might be high due to the strong negative words (e.g., no, doesn't). Neg3 stands out because it resulted in a high negative score in spite of its reading positively as a whole comment; this was due to the negative word usage (e.g., strain, doesn't want to). But it also has a similar neutral score and positive score. We will discuss this misinterpretation of context later in this section.

Comments Pos1-Pos5 in Table 8 are five examples of comments which scored with the highest positive sentiment. Pos1 and Pos2 are relatively short reasoning with positive sentiment, so they both scored 1.00 in positive sentiment. Pos5 also showed shorter reasoning with positive words. Considering that Pos 1, 2, and Pos5 are reasoning from different students. The provided examples uphold the idea that students tend to provide less reasoning when selecting an option.

From the score distribution of positive sentiment and negative sentiment (see Fig. 6), we also found evidence that positive sentiment scores were more widely distributed than negative sentiment. Most of the negative sentiment was found in the reasoning related to the options not selected, while most positive sentiment was found in the reasoning for selected options. The high peak for negative sentiment found in this graph is likely due to the larger number of statements which reasoned for 'selecting an option' over 'not selecting an option.' However, the distribution of the two sentiments indicates that positive sentiment clearly has a larger tail compared to negative sentiment. Specifically, the negative sentiment score had a maximum value around 0.6 while the positive sentiment score had several values higher than 0.6. Taken together, these show that students typically used words with stronger positive sentiments when providing reasoning for the option they claimed was better.

As mentioned earlier, it was also notable that some of the analysis performed by the computer did not capture the context appropriately. Although human researchers can infer students' intention underlying the text, sentiment analysis only discerns word-by-word (or partially syntactic) sentiment in the meaning (see Table 5). For instance,

Table 4 Example preprocessing of students' reasoning into normalized form

Original text	→	Remove punctuation	→	Remove stop words	→	Lemmatized text
It looks like a more comfortable backpack	→	It looks like a more comfortable backpack	→	looks like comfortable backpack	→	look like comfortable backpack
It looks like it has more pockets and space, with the same comfort level	→	It looks like it has more pockets and spaces with the same comfort level	→	looks like pockets space comfort level	→	look like pocket space comfort level
Bags don't close!	→	Bags dont close	→	bags dont close	→	bag dont close

in Table 9, the first comment (Contextfail1) has a low compound score and high negative score due to words such as 'cartoony' and 'problem,' that would generally be interpreted as negatives. Another example, Contextfail2 showed 1.00 neutral sentiment score, but reading the comment actually reveals positive reasoning for the image because it looks 'clearer' and 'more simple.' As a result of preprocessing and limitations of sentiment analysis, some of the comments were not scored appropriately, nor did they reflect student intention.

Frequently used words in the reasoning process of backpack comparison

The most frequently used words in the students' reasoning are depicted below in Table 10, and we describe three insights about students' reasoning. Based on review of the comments, and modified stop words described earlier, we excluded some of the words that were used in relation to the LbE experience instead of the content of the options being shown (e.g., show, picture, image, option).

First, most comments made could be interpreted in relation to the design task and original value function of a backpack—to carry (S1, NS1) or to hold (S10, NS10) something—regardless of the claim. Other commonly used words, for instance, item (S3, NS8), (every) thing (S5, NS5), stuff (NS6), and more specifically, book (NS3), can be interpreted in light of what is carried. Expanding the function of 'carrying' something, students also considered how many (S8)/multiple (NS4) items the backpack can carry, which also related to the idea of "space" (S9); these words suggest that students considered spaces or number of items the backpack could hold as an important criterion for their decisions.

Second, usage or user perspective was considered an important criterion for reasonings, both to select an option or not to select an option. In the reasoning of selected options, how easy (S7) or well (S6) the backpack could carry stuff was often mentioned. As well as thinking about the backpack itself, students took user perspectives into account. In contrast, the reasoning for not selected options showed that students commented on the backpack used words like 'heavy (NS2)' and/or 'hard (to do something) (NS7)', showing additional instances of student concern for the possible user experiences of those with the backpack.

Third, when providing reasonings for the selected options, students also regarded the design (S3) of the backpack as an important justification. Design may represent a broad term that includes overall reactions to the options and details such as aesthetic features of the backpack. Put simply, students rejected options that explicitly failed to fulfill the very basic functions of a backpack (e.g., fail to carry stuff, cannot hold multiple item). Yet, the

Table 5 Exemplary results of sentiment analysis

#	Text	Code	Neg	Neu	Pos	Com
Ex1	Man carrying everything on his back has a lot of **pain**	1	0.45	0.55	0.00	−0.51
Ex2	I think it is efficient, less **painful**	0	0.35	0.27	0.38	0.05
Ex3	Purse **irritates** skin inside elbow (…)	1	0.29	0.42	0.30	0.00

Neg: Negative sentiment score, **Neu**: Neutral sentiment score, **Pos**: Positive sentiment score, **Com**: Compound sentiment score

Table 6 Scores from Sentiment Analysis: Negative, Neutral, Positive and Compound

	Neg	Neu	Pos	Com
	M(SD)	*M(SD)*	*M(SD)*	*M(SD)*
Selected (N = 129)	0.01 (0.15)	0.73 (0.29)	0.26 (0.29)	0.22 (0.25)
Not Selected (N = 19)	0.10 (0.28)	0.78 (0.30)	0.07 (0.14)	0.00 (0.28)

Neg: Negative sentiment score, **Neu**: Neutral sentiment score, **Pos**: Positive sentiment score, **Com**: Compound sentiment score

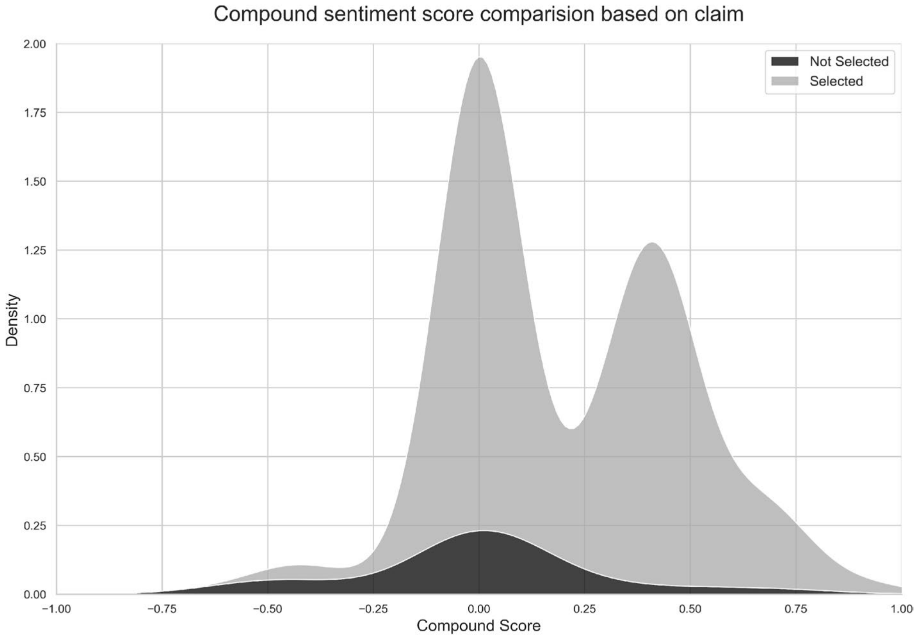

Fig. 5 Compound score distribution of two reasonings: Selected and not selected

analysis showed that if both of the backpack options could meet the criteria of carrying multiple items, students seemed more inclined to make a decision based on the design features present or absent in each of the displayed backpacks.

Table 7 Comments with high negative sentiment score

#	Neg	Neu	Pos	Com	Text	Selected*
Neg1	0.61	0.39	0.00	– 0.48	Poor horse wheels	N
Neg2	0.45	0.55	0.00	– 0.51	The man carrying everything on his back looks to be in a lot of pain	N
Neg3	0.37	0.32	0.21	0.35	Using a Lama to carry you belongings is a good idea because it puts no strain on yourself and who doesn't want to walk with a lama	Y
Neg4	0.24	0.76	0.00	– 0.44	Carrying everything box make arms tired especially carrying every day	N
Neg5	0.23	0.45	0.33	0.22	I feel like the left one is better because what if you don't have kids. Then the backpack has no value	N

Selected* means if it is a reasoning about selected option. **Neg**: Negative sentiment score, **Neu**: Neutral sentiment score, **Pos**: Positive sentiment score, **Com**: Compound sentiment score

Table 8 Comments with high positive sentiment score

#	Neg	Neu	Pos	Com	Text	Selected*
Pos1	0.00	0.00	1.00	0.42	It looks more efficient	Y
Pos2	0.00	0.00	1.00	0.34	Looks more secure	Y
Pos3	0.00	0.12	0.88	0.93	Helps improve confidence and positivity to have a new backpack with a special message on it	Y
Pos4	0.00	0.18	0.82	0.58	That looks cool and that looks cool to the baby	Y
Pos5	0.00	0.26	0.74	0.69	Better material easier to carry	Y

Selected* means if it is a reasoning about selected option. **Neg**: Negative sentiment score, **Neu**: Neutral sentiment score, **Pos**: Positive sentiment score, **Com**: Compound sentiment score

Content analysis of comments in the context of engineering design thinking

For further elaboration and inference, our research team performed qualitative content analysis (Hsieh & Shannon, 2005) based on the previous computer-aided content analysis. As mentioned earlier, computer-aided content analysis may not always capture context or rhetorical nuance of the comments (Neuendorf, 2017; Su et al., 2017). Previous sections also demonstrated this problem (Table 9). Considering that students' comments were situated in the engineering design course, researchers qualitatively analyzed the comments through this lens of design thinking. This stage provided an interpretation of the underlying context of the reasoning and when students made a claim about selecting or not selecting backpacks.

Empathizing by identifying themselves as users

When taking a look at the word frequency, we found students treated themselves as the end-user of the items displayed (e.g., me, my, I). When taking a closer look, we found that some of the students (N = 5) tended to specifically articulate their opinions from this user perspective. Examples from different students are as below.

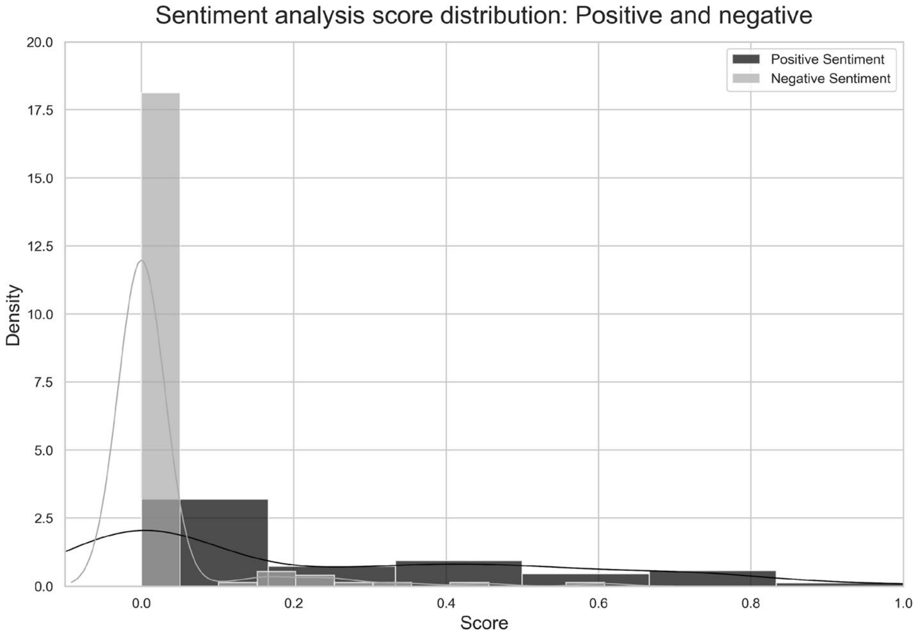

Fig. 6 Sentiment analysis score distribution of positive and negative sentiment

> *"I do not like the clear backpack because you can see everything inside someone's bag, including valuable items such as phones."*
> *"I like backpacks that you can put alot of stuff in."*
> *"I have issues with my backpack being too heavy and not being able to carry a lot of stuff."*
> *"I prefer using bags as opposed to carrying stuff in boxes."*
> *"The cameras backpack is a very good idea because it has snug padding to keep all your gear safe (i need to get myself one of those)."*

In design thinking, a designer's ability to empathize with the users is critical to developing an understanding of the problem. Our analysis showed that students could easily imagine themselves in the user environment demonstrating empathy toward users.

Defining user's needs and problems

As seen in word frequency analysis, students considered users' needs and problems when making a decision. In most cases, students considered four main areas concerning backpacks: the convenience of use (e.g., more secure, enough spaces), characteristics and features (e.g., charge your laptop, carry and move around), specific occasion (everyday use), and fulfilling specific user needs (e.g., people with no kids, carry a baby, walking a lot). Here are representative examples from the student comments:

> *"I think that it is easier to carry a baby than a giant roll of materials."*
> *"Rolling backpacks are very useful if you are walking a lot."*

Table 9 Examples of sentiment failed to catch the sentiment and underlying context

#	Neg	Neu	Pos	Com	Text	Selected*
Contextfail1	0.25	0.75	0.00	−0.44	This image is also more appealing with the cartoony style of presentation, and showing a relatable problem	Y
Contextfail2	0.00	1.00	0.00	0.00	The image seems clearer and more simple. It also shows an additional use for the backpack	Y

Selected* means if it is a reasoning about selected option. **Neg**: Negative sentiment score, **Neu**: Neutral sentiment score, **Pos**: Positive sentiment score, **Com**: Compound sentiment score

Table 10 Frequently used words in reasoning, based on claim

Reasoning of selected option			Reasoning of not selected option		
Selected (S)#	Word	Count	Not Selected (NS) #	Word	Count
S1	Carry	27	NS1	Carry	6
S2	Help	15	NS2	Heavy	5
S3	Design	15	NS3	Book	5
S4	Item	14	NS4	Multiple	3
S5	Thing	12	NS5	Everything	3
S6	Well	12	NS6	Stuff	2
S7	Easy	11	NS7	Hard	2
S8	Many	9	NS8	Item	2
S9	Space	7	NS9	Number	2
S10	Hold	7	NS10	Hold	2

> *"I feel like the left one is better because what if you dont have kids. Than the backpack has no value."*
> *"If you are more of a physical person this would be better."*
> *"Helps improve a lightweight backpack for everyday use."*
> *"A small backpack with a limited carrying capacity but is useful for carrying around art supplies."*
> *"I have chosen A because your arms can get tired of holding your baby so you can hold him using your backpack and you also have space to put your things... cause its a backpack."*

It is meaningful that students made connections to real-world, authentic problems as well as to users with different and specific needs. When students presume themselves as a user, they might specifically mention their own needs. Similarly, when they took possible users (e.g., as parents, physical person, someone carrying art supplies) into consideration, they tend to elaborate on what each specific user is in need. These types of comments reiterate students' empathy while solving problems as designers.

Aesthetic features or ideation to students' own design products

The word 'design' was only used for the reasoning about the selected option. When looking at the comments including this term, students used 'design' for two purposes: to refer to aesthetics or in reference to their own design process. First, in many reasonings the word design was a criterion regarding an option's aesthetically pleasing qualities. It was notable, however, that there were only two students regularly bringing this aesthetic rationale into their decision-making.

"Much better design"
"Looking better design"
"Nicely designed and is more appealing to buyers"
"Has an organized design"

Secondly, students adopted the word 'design' in relation to a plan for their own design model. When provided images encompassing inspirational sketches, actual images of backpack, and up-to-date features, students utilized the images for ideation and commented on how it could inspire them to design a new backpack:

"Helps show that my design should be able to hold heavy items without breaking."
"Mine should have many pockets and be organized."
"This image is more humorous and grabs the viewer's attention better."
"Helps me show that straps could be involved in my design."

However, it is also true that students compared and discussed the quality of the options, rather than ideating from the provided options. Many of the student reasonings simply judged the function, design, and quality of backpack. In other words, in spite of a holistic prompt and suggestion to base their choice on inspiration for iteration and though the comments provided a clear reason for selecting one option over another, it may not have lead them to new inspiration for their own designs.

"It can hold more things"
"I have chosen B because the backpack looks nice"
"A multi-purpose bag is a good idea"
"It can hold larger items"

For instance, the comment 'It can hold more things' clearly showed why the student preferred the selected option. However, students reasoning here may not extend to their own ideation when designing (i.e., realizing that they also need to design a bag which can hold more things). Similarly, 'I have chosen B because the backpack looks nice' implies the student valued aesthetic features of the backpack. Even though they articulated the benefits of the selected option over the not selected one, we cannot conclude whether students considered these aspects when they designed their backpack. Hence, we cannot conclude how the judgment process impacted the design brainstorming of students.

Learning transfer

With respect to the LbE experience preparing students for their upcoming problem solving, it is also notable that some of the reasoning showed signs of failed learning transfer.

Teachers included backpack designs and other images that were intended to stretch the creativity of the students' designs; however, many students could not make a connection between the images—for example, a possum carrying her family (Fig. 7) —and what would be helpful when designing a backpack. Since it was hard to understand the intention of the image being there, three students who were provided images of possum reported confusion in their reasoning. Several other reasonings showed students' failure in understanding the intention of pictures or in their ability to abstract the images to the design process.

> *"Why is this relevant – hamster"*
> *"The other is a literal animal"*

As a result, when presented with these images, the students selected the other option without much evidence supporting the claim. In general, these examples of concrete thinking imply how students' reasoning can be affected by the design and orientation of LbE, including the holistic statement and examples presented. When integrating LbE, this means that ultimately such students were not prepared to recognize opportunities for transfer of learning.

Discussion

This study incorporated content analysis through computer-aided methods (i.e., sentiment analysis and word frequency) and human-based methods altogether to investigate students' reasoning used in LbE. Findings from sentiment analysis indicated that students used words with more positive sentiment when providing reasoning for the option they selected. However, they used words with more negative sentiment when providing reasoning for the option they did not select. In the student reasonings for selected options, students expressed more biased, positive sentiment with less neutral expression and used shorter reasoning. This may be due to the lack of critical thinking or reflection or may result from failing to articulate their thinking in comments in spite of in-depth reasoning. Thus, further educational intervention is required to facilitate logical, detailed, design-thinking-based reasoning, which eventually can be provided in comments.

Results from word frequency analysis showed how students perceive the importance of the design features they need to work on. Original functions of backpack, user convenience, and aesthetic design features were considered as crucial. We found aesthetic features were considered more often when selecting an option, and less considered when not selecting an option. Qualitative analysis of these comments found a few of students just briefly used the term 'design' to describe their intuitive impression or preference toward the option (e.g., a better looking design). Though these comments imply that students cared about aesthetic features when selecting a design, the reasoning itself is still vague.

However, our qualitative analysis indicated that not all students are taking all features into consideration when making a judgment. This is in line with past research that suggests beginners tend to reason with a single explanation (Sandoval & Millwood, 2005). In other words, some of the students used just one or two criteria in their reasoning. For instance, there was one student who only cared about the function of backpack design and failed to incorporate other features making a more efficient backpack. Hence, further intervention such as activities sharing criteria with others can be considered when implementing the LbE into an educational setting. This finding also reinforces the importance of a class-wide discussion following LbE comparison making, as an opportunity to synthesize this

Fig. 7 Picture of possum used to inspire students "©Evelyn / Adobe Stock #282,517,042"

information about what criteria matter in this design context and help students to recognize and deal with interconnectivity and trade-offs as an important part of design decision making (Crismond & Adams, 2012; Goel & Pirolli, 1992).

Finally, content analysis from the researchers found students tend to (1) identify themselves as users, (2) consider user perspective and context, (3) use design as an aesthetic criterion or to ideate their design. When students identified themselves as users, they could emphasize better to users but also had a limitation of only reflecting their own perspective instead of considering broader target users. Furthermore, without the broader perspective, students considered limited user contexts. When requested to use options to ideate a new backpack design, students could incorporate design features successfully through their reasoning. However, there were a few signals of the failure of learning transfer, as well. Effective and explicit introduction of the holistic statement (preparation for the LbE comparisons) may help prevent confusion in students, which leads to the failure of learning transfer. Additional activities helping students share their reasoning process with peers or the entire classroom (debrief on the LbE comparisons) may help students think more broadly and critically.

Suggestion for teacher educators

Exploring the contents of students' reasoning in LbE provided insights into how teachers and STEM design educators can use ACJ as a tool to inform meaningful learning. First, teachers should introduce the LbE with detail and be explicit about the learning intentions. Teachers should present the goal of LbE (e.g., select an option better design, which inspires your own design), contents of judgment (e.g., backpack), with meaningful criteria of judging backpack design (e.g., function, user perspective, creative, aesthetic features) before LbE. This should be well-presented and summarized in a holistic statement the students will see in every judgment. After this, providing modeling of LbE can be helpful. For instance, actually modeling a couple of judgments of the sample work (e.g., previous students' artifact) is a viable and effective option to help students see design reasoning in action.

Without explicit and specific guidance, students may simply select options that are visually pleasing (e.g., looks better, has a cool design) or based on the preference without any reasoning (e.g., It seems more appealing, it looks more efficient), or make poor selections when they fail to understand (e.g., what is this?). Students tendency for these types of selections were supported by evidence of short, positive reasoning without specific evidence in the sentiment analysis. Studies (McNeill & Krajcik, 2008; Ping et al., 2020; Slavit et al., 2021) view reasoning as an essential process of claims based on evidence, and highlight the importance of explicit teaching. This is especially the case when disciplinary content and practices become more explicit, and the claims and reasoning becomes more centered on a specific content area with epistemological shifts (Slavit et al., 2021). Thus, when specific and explicit context provided to the students with more details, student-thinking in an articulated manner is expected to be exhibited.

Second, post-LbE activities such as teacher-led or peer-led discussion based on actual judgments can be helpful for leveraging understanding. While some of the students could articulate their reasoning in a substantially diverse way, with a good understanding within design thinking context, other students fell short. When given an opportunity to follow up on comparisons, students can think about how to use preferred features they have seen in the ideation process, while preventing inclusion of the less preferred features. Also, students have a chance to clarify understanding of the options with peer discussion. For instance, as presented earlier, some of the students could not successfully understand what the possum picture has to offer backpack design (Fig. 7). They could not imagine the transfer applications and may have assumed it was an attempt at testing engagement. Not only can a discussion facilitate expanding the characteristics and features of good design, but it can also help students articulate why they think it looks better. In a more public setting (e.g., small group, entire classroom discussion) which requires interaction and participation, students need to provide explanation with more details, which eventually promotes articulation of their knowledge (Collins et al., 2018).

Third, post-LbE discussion may help students think from others' perspective, including users. In some cases, students have contrasting opinions toward the same option (e.g., a simplistic backpack that can carry a good amount of beachwear vs. I do not like the clear backpack because you can see everything inside someone's bag, including valuable items such as phones). Discussion on these subjects help students foster a design mindset with openness to diverse opinion (Brenner et al., 2016; Ekman & Ekman, 2009). Though students in this study empathized with users, they may be biased if they stick to their own perspective and identify only themselves as users. In this sense, discussion after LbE can help balance empathy and open mindset toward a broad user group. Rauth et al., (2010) underscored the importance of the involvement of all perspectives when designing. They see each team and individual developing their own process as they work on a problem and adapting and adding to the solution as the essence of design thinking. Being mindful of the opinion of others can foster creative thinking among students, allowing them 'thinking out of the box' from one single individual's perspective.

Limitation

Several limitations in this study should be acknowledged. Sample size of the current study is small compared to other studies using machine learning and natural language processing. Though qualitative content analysis compensates for the sample size, an imbalanced dataset with small samples may not be generalizable to all contexts. Some aspects of the LbE session were inconsistent in their presentation to students. There were some technical glitches students faced while using the software, thus they could not see options clearly. Additionally, size of the images was not consistent throughout the judgment process. Therefore, students may be inclined to select larger images with high resolution unless the resolution was too high, and the image was slow to load. These challenges may have affected students' judgment or explanations in ways that are not associated with the content of the options presented.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical approval The current study was supported by funding from National Science Foundation (NSF), Division of Research on Learning (DRL), Project no. 2101235. Entire research process in the current study involving human participants were in compliance with the ethical standards (i.e., IRB) of the institutions Purdue University (IRB-2021–543). Brigham Young University (R0303035), and University of Georgia (PROJECT 00004032). and school district RRB (file 2021–015). Informed consent was obtained from all individual participants in the study. All the software (RM Compare, Python, Python libraries) has an appropriate permission for the use.

References

Akhtyamova, L., Alexandrov, M., Cardiff, J., & Koshulko, O. (2018). Opinion mining on small and noisy samples of health-related texts. *Conference on Computer Science and Information Technologies*, 379–390.

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Bartholomew, S. R. (2021). Investigating the impacts of differentiated stimulus materials in a learning by evaluating activity. *Mississippi Valley Technology Teacher Education Conference*.

Bartholomew, S. R., & Yauney, J. (2022). The impact of differentiated stimulus materials in learning by evaluating. *Pupils' Attitudes Towards Technology Annual Conference*.

Bartholomew, S. R., & Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *International Journal of Technology and Design Education, 32*(1), 1159–1190.

Bartholomew, S. R., Mentzer, N., Jones, M., Sherman, D., & Baniya, S. (2020). Learning by evaluating (LbE) through adaptive comparative judgment. *International Journal of Technology and Design Education, 32*(2), 1–15.

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., & Reeve, E. M. (2018a). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment, 23*(2), 85–101.

Bartholomew, S. R., Strimel, G. J., & Jackson, A. (2018b). A comparison of traditional and adaptive comparative judgment assessment techniques for freshmen engineering design projects. *International Journal of Engineering Education, 34*(1), 20–33.

Bartholomew, S. R., 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

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Brenner, W., Uebernickel, F., & Abrell, T. (2016). Design thinking as mindset, process, and toolbox. In W. Brenner & F. Uebernickel (Eds.), *Design thinking for innovation* (pp. 3–21). Springer.

Canty, D., Seery, N., Hartell, E., & Doyle, A. (2017). Integrating peer assessment in technology education through adaptive comparative judgment. *PATT34 Technology & Engineering Education–Fostering the Creativity of Youth Around The Globe, Millersville University, Pennsylvania, USA*, 10–14.

Collins, A., Brown, J. S., Newman, S. E., & Resnick, R. (2018). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In *Knowing, learning, and instruction* (pp. 453–494). Routledge.

Cowan, J. (2010). Developing the ability for making evaluative judgements. *Teaching in Higher Education, 15*(3), 323–334.

Crismond, D. P., & Adams, R. S. (2012). The informed design teaching & learning matrix. *Journal of Engineering Education-Washington, 101*(4), 738–747.

Daly, S. R., Adams, R. S., & Bodner, G. M. (2012). What does it mean to design? A qualitative investigation of design professionals' experiences. *Journal of Engineering Education, 101*(2), 187–219.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, 519–528.

Dewit, I., Rohaert, S., & Corradi, D. (2021). How can comparative judgement become an effective means toward providing clear formative feedback to students to improve their learning process during their product-service-system design project? *Design and Technology Education, 26*(3), 276–293.

Dey, L., & Haque, S. M. (2008). Opinion mining from noisy text data. *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, 83–90.

Dey, L., & Haque, S. M. (2009). Studying the effects of noisy text on text mining applications. *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, 107–114.

DeKalb County School District. (2021). *DeKalb County School District Demographics (2021–2022 School Year)*. https://www.dekalbschoolsga.org/about/

Van Eemeren, F. H., Grootendorst, R., & Kruiger, T. (2019). Handbook of argumentation theory. In *Handbook of Argumentation Theory*. De Gruyter Mouton.

Ekman, S., & Ekman, A. (2009). Designing an entrepreneurial mindset in engineering and management. *DS 58–9: Proceedings of ICED 09, the 17th International Conference on Engineering Design, Vol. 9, Human Behavior in Design, Palo Alto, CA, USA, 24.-27.08. 2009*, 179–190.

Fielding-Wells, J. (2016). Mathematics Is Just 1+ 1= 2, What Is There to Argue About?: Developing a Framework for Argument-Based Mathematical Inquiry. *Mathematics Education Research Group of Australasia*.

Forman, J., & Damschroder, L. (2007). Qualitative content analysis. In L. Jacoby & L. A. Siminoff (Eds.), *Empirical Methods for Bioethics: A Primer* (Vol. 11, pp. 39–62). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3709(07)11003-7

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior, 52*, 315–325.

Gillies, R. M. (2020). *Inquiry-based science education*. CRC Press.

Goel, V., & Pirolli, P. (1992). The structure of design problem spaces. *Cognitive Science, 16*(3), 395–429.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Hokanson, B. (2012). The design critique as a model for distributed learning. In L. Moller & J. Huett (Eds.), *The next generation of distance education* (pp. 71–83). Springer.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288.

Jackson, A., Bartholomew, S., & Mentzer, N. (2022, November 17). *Priming the design process: Activating and characterizing students' critical thinking in design* [Paper presentation]. 108th Mississippi Valley Technology Teacher Education Conference and the 60th Southeastern Technology Education Conference, Nashville, TN.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93–101.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication, 58*(2), 258–279.

McNeill, K. L., Katsh-Singer, R., & Pelletier, P. (2015). Assessing Science Practices: Moving your class along a continuum. *Science Scope*, *39*(4), 21–28. https://www.proquest.com/scholarly-journals/assessing-science-practices-moving-your-class/docview/1735273036/se-2?accountid=13360

McNeill, K. L., & Krajcik, J. (2008). Inquiry and scientific explanations: Helping students use evidence and reasoning. In J. Luft, R. Bell, & J. Gess-Newsome (Eds.), *Science as inquiry in the secondary setting* (pp. 121–134). National Science Teaching Association.

Mejova, Y. (2009). *Sentiment analysis: An overview*.

Mentzer, N., Lee, W., & Bartholomew, S. R. (2021). Examining the validity of adaptive comparative judgment for peer evaluation in a design thinking course. *Frontiers in Education, 6*, 1–15.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551.

Neuendorf, K. A. (2017). The content analysis guidebook. In *The content analysis guidebook* (2nd edition.). SAGE Publications, Inc.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102–122.

Perkins, D. N., Crismond, D., Simmons, R., & Unger, C. (1995). *Inside understanding*. Oxford University Press.

Ping, I. L. L., Halim, L., & Osman, K. (2020). Explicit teaching of scientific argumentation as an approach in developing argumentation skills, science process skills and biology understanding. *Journal of Baltic Science Education, 19*(2), 276–288.

Plisson, J., Lavrac, N., & Mladenic, D. (2004). A rule based approach to word lemmatization. *Proceedings of IS, 3*, 83–86.

Pollitt, A., & Whitehouse, C. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. Center for Education Research and Policy.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2), 143–157.

Rauth, I., Köppen, E., Jobst, B., & Meinel, C. (2010). Design thinking: An educational model towards creative confidence. *DS 66–2: Proceedings of the 1st International Conference on Design Creativity (ICDC 2010).*

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment and Evaluation in Higher Education, 35*(5), 535–550.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23–55.

Sebeok, T. A., & Zeps, V. J. (1958). An analysis of structured content, with application of electronic computer research, in psycholinguistics. *Language and Speech, 1*(3), 181–193.

Slavit, D., Grace, E., & Lesseig, K. (2021). Student ways of thinking in STEM contexts: A focus on claim making and reasoning. *School Science and Mathematics, 121*(8), 466–480.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998b). *The use of self-, peer-and co-assessment in higher education*.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1998a). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research, 1*(3), 293–319.

NGSS Standards. (2013). *Next generation science standards: For states, by states*. The National Academies Press DC.

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75.

Strimel, G. J., Bartholomew, S. R., Purzer, S., Zhang, L., & Ruesch, E. Y. (2021). Informing engineering design through adaptive comparative judgment. *European Journal of Engineering Education, 46*(2), 227–246.

Su, L.Y.-F., Cacciatore, M. A., Liang, X., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2017). Analyzing public sentiments online: Combining human- and computer-based content analysis. *Information, Communication & Society, 20*(3), 406–427.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26*(3), 249–269.

Thurstone, L. L. (1941). *Factorial studies of intelligence*. University of Chicago Press.

Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*(1), 47–58.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Wilson-Lopez, A., Strong, A. R., Hartman, C. M., Garlick, J., Washburn, K. H., Minichiello, A., Weingart, S., & Acosta-Feliz, J. (2020). A systematic review of argumentation related to the engineering-designed world. *Journal of Engineering Education, 109*(2), 281–306.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1253.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.