International Journal of Technology and Design Education (2024) 34:1561–1577
https://doi.org/10.1007/s10798-023-09871-5

Investigating the impacts of differentiated stimulus materials in a learning by evaluating activity

Scott Bartholomew[1] · **Jessica Yauney**[1] · **Nathan Mentzer**[2] · **Scott Thorne**[2]

Accepted: 6 December 2023 / Published online: 5 January 2024

Abstract
Classroom research has demonstrated the capacity for significantly influencing student learning by engaging students in evaluation of previously submitted work as an intentional priming exercise for learning; we call this experience *Learning by Evaluating* (LbE). Expanding on current LbE research, we set forth to investigate the impact on student learning by intentionally differing the quality of examples evaluated by the students using adaptive comparative judgement. In this research, university design students (N = 468 students) were randomly assigned to one of three treatment groups; while each group evaluated previously collected student work as an LbE priming activity, the work evaluated by each group differed in quality. Using a three-group experimental design, one group of students only evaluated high quality examples, the second only evaluated low quality examples, and the third group of students evaluated a set of mixed-quality examples of the assignment they were about to work on. Following these LbE priming evaluations, students completed the assigned work and then their projects were evaluated to determine if there was a difference between student performance by treatment condition. Additional qualitative analysis was completed on student LbE rationales to explore similarities and differences in student cognitive judgments based on intervention grouping. No significant difference was found between the groups in terms of achievement, but several differences in group judgement approach were identified and future areas needing investigation were highlighted.

Keywords Adaptive comparative judgment · Assessment · Design-thinking · Learning by evaluating · Peer assessment

✉ Scott Bartholomew
scottbartholomew@byu.edu

Jessica Yauney
jessica.yauney@gmail.com

Nathan Mentzer
nmentzer@purdue.edu

Scott Thorne
sthorne@purdue.edu

[1] Brigham Young University, Provo, USA

[2] Purdue University, West Lafayette, USA

Introduction

While K-12 classrooms engage with learning and assessing daily, assessment is often viewed as the role of the teacher, having limited effect on student learning through teacher choice to revisit or expand on earlier material (Johnson et al., 2019). In general, assessment practices have improved over time (e.g., greater access to technology for facilitating assessment) (Robertson et al., 2019), but relatively little has changed in terms of students' participation in assessment processes—for the most part, in a linear sequence, students submit work, teachers evaluate this work, assign grades, and then the teacher moves to a new topic. This formulaic approach to assessment often coincides with assessment signaling *the end* of student learning as opposed to a key step in the learning process (Bartholomew et al., 2020). Recent work with evaluation—a key element of assessment traditionally engaged in by the teacher—has demonstrated the potential of these evaluation activities to play a much larger role in students' learning. Specifically, as students are engaged in evaluation activities and then encouraged to revisit/review/revise their own work, their learning has been significantly positively impacted (Sherman et al., 2022; Bartholomew, Strimel, & Yoshikawa, 2019). Research into this approach, referred to as "Learning by Evaluating" (LbE), has highlighted this potential; instead of viewing peer-assessment as a task meant to benefit their peers, students can intentionally engage in the evaluation process as a beginning step in their own learning and comprehension. In this vein, LbE has demonstrated that as students engage with exemplar work, they exercise higher order thinking skills (e.g., evaluation and analysis) that can help strengthen their own understanding of the task, the associated requirements, and the applicable skills, aptitudes, and approaches.

Literature review

Given its rise from assessment-focused research (e.g., comparative judgment and adaptive comparative judgment), it is not surprising that questions around LbE remain and connections to existing bodies of literature and research are not fully developed. However, early work suggests that the theoretical basis for Learning by Evaluating (LbE) is informed by multiple areas of study including cognitive apprenticeship and Bloom's taxonomy of learning.

Learning by evaluating

Following research into improving design education, several researchers (e.g., those in the UK (Kimbell), Ireland (Seery & Canty), Sweden (Hartell) and the USA (Bartholomew)) recognized the potential for utilizing evaluation as a learning tool for students rather than simply as an assessment approach for teachers. Specifically, Bartholomew & Yauney (2022) coined the term *Learning by Evaluating* (LbE) to describe a process wherein students use, view, and evaluate multiple pairs of work in an ACJ setting *prior* to engaging in the assignment themselves. Several studies have shown positive results in terms of student learning through LbE with implications of this approach to facilitate student learning and growth (Baniya et al., 2019; Bartholomew & Yauney 2022; Bartholomew & Strimel 2019; Bartholomew et al., 2018b; Bartholomew et al., 2018a; Seery & Canty, 2017). Students have specifically called out benefits of this approach such as its ability to help them gain

confidence (Canty, 2012) and improve their own work (Bartholomew et al., 2019). This process has been applied in a variety of fields and has been shown to have positive effects in a myriad of courses such as undergraduate design courses, English, Engineering, and Business (Bartholomew & Jones, 2020).

Cognitive apprenticeship

In an LbE experience, a learner critically evaluates previously submitted work and, during this evaluation experience, engages in several methods inherent in cognitive apprenticeship including reflection, modeling, and articulation (Collins, 2022). During the reflection portion of cognitive apprenticeship, a student is invited to "compare their own thinking processes with those of an expert or other students" (Collins, 2022, p. 1). Likewise, during LbE, students engaged in critical evaluations of example work—these are often completed through pairwise adaptive comparative judgment comparisons and inherent in these evaluations is a comparison of what is displayed (e.g., previous student work) with a student's own ideas, thoughts, plans, and intentions for the same assignment. Bartholomew, Strimel, and Yoshikawa (2018a, 2018b) highlight the benefits of this reflection and comparison process by citing several student's talking about their experience. These students shared:

> *I learned about some things I could do for my [assignment] that would make it better. For example, many people had more graphics and design elements than me, so I added more of these to mine. There was also more information and detail in certain sections which I decided to reflect in my own project (p. 376).*
> *I learned that my peers do a lot of things differently than I do them. For example, I formatted my project to the example, where others did a totally different format (p. 377).*

As Collins, (2022) points out, during the *modeling* phase of cognitive apprenticeship an expert performs a task and students learn through observation. This process is similarly undertaken in LbE as students can see how a task/assignment has been accomplished by their peers *before* starting on the same task themselves. While the argument may be made that peer modeling lacks an "expert" to perform the modeling as is usually seen in cognitive apprenticeship, research has demonstrated the potential for learning through the observation of both high- and low-quality examples (Schwartz et al., 2011).

In addition to connections with modelling, Kimbell, (2018) highlighted another benefit of LbE which relates directly to the *articulation* methods contained in cognitive apprenticeship. He noted (p. 185):

> *We know that, in design terms, the act of expression [e.g., articulation] pushes ideas forward. So too with this [LbE/ACJ] discourse, the act of [comparison and evaluation] begins to crystallize the construct for them. It makes a vague and intangible construct into something a bit more substantive*

Similarly, Bartholomew et al., (2018a, 2018b) conducted research with middle school students engaged in an LbE experience and cited one student who noted that the process of articulating the evaluation decisions during the review of peer work was beneficial to their learning. Specifically, this student shared that:

> *I think [articulating rationale for evaluation decisions] helped me because people gave me direct information on what information I need to use or put in (p. 381).*

Considering these connections between cognitive apprenticeship and the LbE approach, we posit that cognitive apprenticeship may provide both a rationale for, and a theoretical basis from which to build our understanding of LbE. Specifically, we see cognitive apprenticeship as a means for describing both why and how LbE may impact student learning.

Bloom's taxonomy of learning

In addition to a foundation in cognitive apprenticeship, elements of LbE employ concepts from Bloom's Taxonomy of learning (1956). Specifically, Bloom's taxonomy notes several categories of learner experience ranging from "remember" to "evaluation." Each taxonomy lies along a continuum, from simple to complex, of potential student learning skills. At the onset of an LbE learning experience, students compare examples of previous student work (analyze) and evaluate which is better (evaluation); in this way they practice two of the highest skills along Bloom's learning taxonomy. Further, the act of articulating an evaluation decision rests firmly in the evaluation portion of Bloom's taxonomy. In this way, and like other research (e.g., contrasting cases—see Schwartz, et al., 2011), LbE rests on the premise that specifically engaging students in higher-order thinking skills (evaluation) that lead to great learning gains (Collins, 2014) as opposed to other learning activities with an emphasis on lower-order methods (e.g., remembering).

Adaptive comparative judgement

Although LbE is not dependent on Adaptive comparative judgment (ACJ), the research into LbE has largely utilized ACJ as a vehicle through which the LbE is accomplished (Bartholomew & Jones, 2020). By itself, ACJ is a structured method of assessing items (e.g., student work) while making a series of comparative evaluations. In ACJ, an individual—or a group of individuals—view pairs of items and determine, based on an identified criterion, which is better. This process is repeated iteratively with different pairings of items. Though we are using ACJ as a primer where our desired outcome is the process of comparison without regard to which item is better, the process was originally developed to establish a rank order. This approach to ACJ is built on comparative judgment (CJ) which was developed by psychometrician Thurstone, (1927) and later refined by academic Alastair Pollitt, (2004, 2012). Both CJ, and subsequently ACJ, were proposed as an alternative approach to assessment through rubrics or other criterion-based approaches. The original idea of randomly working through a series of paired comparisons (CJ) was later extended by adding an adaptive algorithm to this approach—hence the "A" in ACJ—which served to facilitate faster and more reliable judgment results using automated technology software (Pollitt, 2012; Rangel-Smith & Lynch, 2018). Data resulting from CJ and ACJ includes a rank-order of all items, comparison decision rationale, judgment times, and Rasch-modeling misfit data associated with the process (Bartholomew, 2017; Pollitt, 2004, 2012). Previous research has shown high reliability levels (Baniya et al., 2019, Bramley, 2015; Mentzer, Lee, & Bartholomew, 2021), a simpler assessment process (Kimbell, 2021), and greater ease of integrating assessment feedback from multiple assessors (Bartholomew & Yoshikawa, 2018; Kimbell, 2012) over other more traditional approaches to evaluation. While a complete explanation of ACJ is beyond the scope of this work, further information can be found in the works of Pollitt, (2004, 2012, 2015), Bramley, (2015), and Rangel-Smith and Lynch (2018).

Research question

With various LbE (and associated ACJ) research demonstrating the potential for enhancing student learning, questions around the potential to modify or enhance LbE have risen (Bartholomew et al., 2020; Buckley, Kimbell, & Seery, 2022. Specifically, findings related to increased student achievement following LbE have been accompanied by questions into how the examples evaluated by the students may influence the subsequent learning and performance of students; specifically, Bartholomew & Yauney (2022) questioned the potential for improving LbE by intentionally varying the quality of work presented to students during LbE. If LbE has demonstrated a significant impact on student learning using previous student work of varying quality, can LbE be improved even further through intentionally selected items for evaluation? In response to this call, we determined to investigate this potential. The research question which guided our efforts was:

RQ: What is the influence of item quality, if any, on student learning during LbE?

Based on previous work and findings (e.g., Bartholomew & Yauney 2022), we hypothesized that varying the quality of examples during LbE may influence student learning. Specifically, we hypothesized that students in LbE with only high-quality examples may rise to a "higher standard" which would lead to better educational outcomes—a hypothesis we deemed appropriate based on observations as educators who have used LbE in class, and which aligns with research that shows that high expectations generally lead to an increase in student achievement (Johnston et al., 2019). Relatedly, we also hypothesized that engaging students in LbE with only low-quality examples may lower their expectations for the assignment and subsequently result in them performing worse than their peers. Finally, we hypothesized that engaging students with mixed-quality examples may lead to the highest educational outcomes based on an opportunity to identify strengths of high-quality examples and weaknesses of low-quality examples (e.g., in line with research on contrasting cases by Miksza, 2011 and Schwartz, et al., 2011; the work of Caniglia, 2020 around learning from the mistakes of others; or the work of Kimbell, 2018 suggesting higher levels of learning from work in varying quality).

Methods

To better understand the impact of item quality in LbE we completed a mixed methods study; investigating the results both quantitatively (e.g., which treatment led to the highest quality items?) and qualitatively (how were student experiences in each condition different?) by first organizing three collections of items: (1) high-quality examples, (2) low-quality examples, and (3) mixed-quality examples. These examples were taken from students' submissions for a single assignment centered on the creation of a point-of-view design statement (POV) during 2020. POVs identify a user, their unique need, and an insight meant to guide the designer in producing a potential solution (see Sherman et al., 2022; Wible, 2020; Dam & Siang, 2020; or Mentzer, Lee, & Bartholomew, 2022 for a more-detailed explanation on POVs). The quality of each student generated POV was determined through an ACJ session conducted by course instructors at the conclusion of the semester in which these statements were created (2020). While completing the ACJ evaluation to determine the quality of the items in preparation for this study, the instructors completed

16 rounds of comparative judgments and reached an overall reliability of $r = 0.71$. This suggested strong levels of agreement and the rank order, produced through the instructor ACJ session, was used to separate the 125 student POVs into high-, low-, and mixed-quality groups. Specifically, the student POVs ranked 1–31 were categorized as "high-quality" examples, POVs ranked 95–125 were categorized as "low-quality" examples, and every fourth POVs in the ranking (e.g., 1,5, 9,…117, 121, 125) were categorized as a "mixed-quality" example. The results from this process were a collection of 31 POVs for each of our three treatment groups to evaluate through LbE representing high-, low-, and mixed-quality examples from the 2020 year.

After preparing our three sets of items (high-quality, low-quality, and mixed-quality), we engaged 468 students in our study of LbE during 2021. These students were all enrolled in an introductory undergraduate design course at a large Midwestern University. During this course, students learned about a design process and engaged in several design experiences. Specifically, at the time of our intervention, these students were all working in groups of 3–5 (n = 112 groups) to complete an 8 week design project. We intentionally opted to provide all students enrolled in the course with an LbE experience based on previous research which showed that LbE provided better educational outcomes for students (Bartholomew & Yauney 2022); therefore, this research did not include a traditional "control group" which would have not received any LbE experience in class.

As part of their eight-week design project, each group was assigned the task of creating a POV statement which would serve as a guide for their subsequent design efforts. Prior to engaging in the task of creating the POV statement, all 112 groups of students (N = 468 students) were randomly assigned to one of the treatment conditions (high-, low-, or mixed-quality examples). These treatment condition assignments, which were known only to the research team, were intentionally spread across teacher and course section to mitigate potential differences in findings based on section or teacher; however, to ensure consistency across group-member experience, all students in each design group were assigned to the same treatment condition. Following the assignment of each student to a treatment condition, all students engaged in LbE individually outside of class as part of their preparatory homework for the next class meeting. This LbE was part of their preparation for the assignment where they would write their own group POV and all students were provided login credentials and instructions for completing the LbE through the ACJ software (RMCompare) as part of this preparatory assignment. During their homework, all students were expected to log in and complete the LbE by viewing the assigned examples, making comparative judgments on quality, and typing a rationale for each decision made (e.g., why they chose one POV over another). Each student viewed ~ 12 different examples of POV statements (six pairs)—differentiated in quality by treatment condition—and chose the better of the two displayed in line with the ACJ process (see Fig. 1). All evaluations, typed rationales, and other ACJ-generated data were collected and separated by treatment condition for later analysis.

Following the intervention, students worked in their groups to create POV statements and fulfill the remaining requirements of their 8 week design project. Importantly, aside from the LbE session they experienced in class (which was less than 20 min in length and engaged the students with high-quality items only, low-quality only, and mixed-quality items), all other classroom procedures, routines, schedules, and assignments were held constant across groups. At the conclusion of the 8 week design experience, all student groups submitted their final POV statements, as part of a larger design portfolio, for assessment.

The final 112 POVs from all students (N = 468) were collected and evaluated to investigate the potential for differences in quality by treatment condition. This evaluation

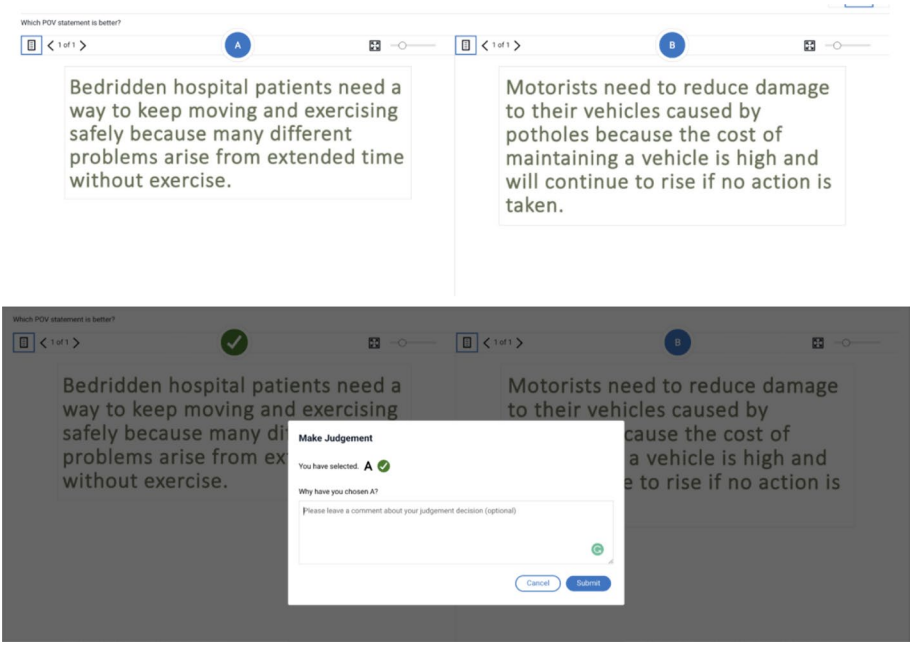

Fig. 1 LbE POV student view

was completed through an ACJ session which was used to determine a rank order of the quality at the conclusion of the 2021 course by students enrolled in the class. Our decision to utilize the students as evaluators during this second ACJ session was intentional—previous research (e.g., Sherman et al., 2022; Strimel, Bartholomew, Purzer, Zhang, & Yoshikawa-Ruesch, 2020) has consistently demonstrated high levels of reliability in student ACJ evaluations and strong correlations with both course instructors and industry professionals (Bartholomew & Jones, 2020). Lastly, given the timing of this evaluation (at the conclusion of an entire semester course on the given topic), we deemed it appropriate to engage these students as evaluators of quality based on their knowledge and experience derived from the course. The resulting ACJ statistics from this final evaluation of all student-created POVs yielded a high reliability ($r = 0.83$) lending credence to our approach and the consistency of judgment decisions made by students. The resulting rank order, and parameter values (a statistic produce by the ACJ software similar to a rank-ordering which shows both order and magnitude of difference between items), and student decision rationale were collected and organized for later analysis.

Once data collection and preparation were completed, ANOVA statistical analyses were performed on the quantitative data (i.e., parameter values) through statistical analysis software (SPSS, v24) to determine what difference, if any, existed between the final student POVs and how the student learning may, or may not, have been impacted by the quality of POVs viewed during their LbE experience. In addition to identifying the potential difference in group achievement, we performed qualitative analyses of the student comments to understand the different experiences of students in the various treatment groups. This analysis consisted of three exploratory analyses completed on students' comments collected during their LbE experience in line with the stated research question.

Qualitative analysis 1

In the first analysis, we chose to investigate the data holistically by analyzing the prevalence of terms contained in the student LbE rationales from the different groups. In our count for terms, we noted synonyms as well as words with the same root (ex. Descriptive and Describes) provided by students to explore trends around student learning during LbE. All comments (2299) made by students while completing the comparisons were analyzed in line with our general research inquiry around the nuances of the LbE experience and followed recommendations by Saldaña, (2015) for attribute coding with frequency counts. In this analysis, the comments from student's evaluations during the LbE exercises were combined and the number of times relevant and related words were used in the decision rationales were calculated. Data were sorted in line with the intervention groups (High-quality POVs, Low Quality POVs, or Mixed-Quality POVs) and general (non-relevant/related) words that did not contribute to the overall meaning of each comment (e.g., "the," "to," or "and") were removed. This frequency list was then put into a table sorted by the frequency of the comments and used as a means of holistically illuminating similarities and differences in the student experience from each of the groups. Further, these data were useful in expanding and triangulating findings derived from the other analyses performed during this effort.

Qualitative analysis 2

Secondly, in line with Saldaña's, (2015) recommendations for attribute coding using thematic categories, and as part of the general inquiry into the potential for influencing student learning through intentionally-varying the quality of items viewed during evaluations, we analyzed the overall sentiment—as opposed to the content—of students' rationales. Specifically, we were interested in how "positively" or "negatively" the student experience may have been impacted as a result of the items the compared during LbE (e.g., did students who only viewed high-quality items have a more "positive" experience in LbE than their peers who viewed only low-quality examples?). For each of the 2299 LbE comments, the student remarks were coded as either purely positive, purely negative, or neutral. Student comments that provided positive feedback generally included statements using words like "good," "better," and "more organized." These were coded as positive while student comments that provided negative feedback—making statements using words like "worse," "more confusing," and "missing" were coded as negative. All rationales that included both positive and negative comments were coded as neutral.

Qualitative analysis 3

Lastly we sought to identify the relationship, if any, between the student rationales they provided during the during the initial LbE intervention and the comments provided by students when evaluating finalized POV statements at the end of the course. For example, we wondered if a student who consistently comments on one element of the POV statement at the beginning was able to translate this focus into superior performance on that element when creating their own POV. In this third analysis, 20 students were randomly selected from each treatment ($k = 60$) (High-Quality POVs, Low-Quality POVs,

or Mixed-Quality POVs) and a thematic analysis of the six rationales they provided during the LbE experience at the beginning was completed in line with recommendations from Baker and Edwards (2012) for qualitative analysis. Then, for each of these students, the rationales given by peers during the evaluation of their POVs at the conclusion was also gathered. The number of peer feedback comments varied both due to the adaptive nature of the software used and because students sometimes commented on only one of the two artifacts being compared but usually included approximately 30 comments. Additional data conditioning was performed to ensure any feedback included in this analysis was specific to the identified student's POV and all elements referenced in the students' comments (both their rationale during the LbE intervention and the rationales provided by peers during the final POV evaluation) were identified. For example, if a student commented "It is a more concise problem that is actionable with more direct purpose" then the themes of being "concise" and "actionable" were identified. Finally, each of these comments was qualitatively analyzed to explore the potential for correlation between the themes in students' LbE rationale and those from the rationale provided by peers during the final POV. See Table 1 and 2 for a full example below.

Findings

Based on previous findings which have shown that students who use LbE have better academic outcomes than those who do not use LbE (Bartholomew & Yauney 2022), we determined to investigate the potential, if any, to influence student learning outcomes by intentionally differing the quality of items evaluated during LbE. Using an ANOVA, we investigated the relationship between student intervention grouping and final POV quality using the parameter values derived from the final ACJ session conducted with final student POVs. Our analysis revealed no statistically significant difference between students who were exposed to high-, low-, or mixed-quality examples ($p = 0.809$). Specifically, the difference between high-quality ($M = 0.13$, $SD = 0.836$), low-quality ($M = -0.07$, $SD = 0.970$) and mixed-quality ($M = -0.04$, $SD = 1.12$) groups

Table 1 Example of analysis completed regarding student themes

Initial student comments	Themes identified
They started out with a long-winded POV statement but eventually came out with a refined version that I think fits very well	Focus
I think option B is more actionable while maintaining focus and direction. Both statements have clear stakeholders, needs, and insights	Actionable, Focus, Direction, Stakeholder, Need, Insight
While A is a longer POV statement, I think it is still more defined and focused	Focus
This one melds the components of a strong POV statement slightly better than the other	Components
I like the insight of this one. It is more surprising	Insight
This one appeals to me. I want to see how it would work. The videos we watched in class explaining POV statements said that a good statement is attractive and option B is attractive to me	Attractive

All samples of feedback are provided by a single student completing an LbE session

Table 2 Samples of student feedback received in second LbE session

Feedback received on student's project	Themes identified	Application made by student following LbE?
I choose this one because although 'it is **short**	Focus	Yes
More **specific** to an issue and more **actionable**	Focus, actionable	Yes
This one is more **actionable** and has a better **stakeholder**	Actionable, stakeholder	Yes
Option A is **too broad** and doesn't provide the same **insight**	Focus, insight	No
The **stakeholder** is clearly defined	Stakeholder	Yes
Option B is **ambiguous**, problem is not defined well enough for a solution to be created	Clarity	No
The POV statement is more **detail**ed than B	Direction	Yes
I chose B because A went into more detail and had **too much** in its statement	Focus	No
It seems easier to **identify all the parts** needed in a good POV statement	Components	Yes
Has all the **requirement components** in order to be a good POV statement	Components	Yes
PoV A seems to have a properly identified group of **stakeholders/users** compared to PoV B	Stakeholder	Yes
It has a **user, a need,** and an **insight** with much fewer words than A	Stakeholder, need, insight, focus	Yes
Very simple, **to the point** without too much detail	Focus	Yes
Has all three descriptively	Components	Yes
better **focus**	Focus	Yes
This statement is clear and **concise**	Clarity, focus	Yes
POV B is too **vague** and does not offer a solution or any **insight** for the design team	Focus, insight	No
B is **unclear**	Clarity	No

This table documents feedback provided to the same student from Table 2. Similar themes are identified, and an indication of if the student applied themes identified in the first LbE session to the one they created in the second LbE is indicated with a 'Yes' or 'No'

Bolded words represent key words used to identify the themes

were not significant either overall or between each of the groups. Further, we noted that each group had a similar number of items ranked in the top and bottom quartiles.

Following this quantitative analysis, and in line with our stated research question, we next investigated the potentially different experiences of students in each of the treatment groups through the qualitative data provided by students' comments on the 12 example POVs (six comparisons) they were shown. This qualitative analysis was completed in three phases – each of these, and the associated findings, will be discussed in turn.

The first analysis of open-ended comments consisted of analyzing the words included in the student rationale provided during the LbE intervention. This was done to further investigate our research question around what influence, if any, the differences in item quality viewed during LbE may have on students learning. Specifically, we investigated the frequency with which these different terms appeared in the LbE rationales for students within each of the groups (High-Quality POVs, Low-Quality POVs, or Mixed-Quality POVs). This analysis of word frequency (see Table 3) showed that the vocabulary specific to elements of POV statements were the most common with "*need*", "*insight*", "*stakeholder*", and "*user*" each appearing more than 250 times in students' comments. In addition to these words, there were many instances of descriptors of the writing in the examples with terms such as "*clear*", "*specific*", "*focus*", and "*detail*" each appearing more than 150 times in students' comments. Using a chi-square test to

Table 3 Word frequency in student comments by group

Term	High-quality group	Mixed-quality group	Low-quality group	Total
Need	163	195	195	553
Clear	130	147	151	428
Insight	139	153	127	419
Stakeholder	118	146	120	384
Specific	115	95	86	296
User	108	67	104	279
Solution	70	81	103	254
Focus	59	77	53	189
Action	67	65	57	189
Detail	69	59	43	171
Format	37	53	58	148
Define	33	47	39	119
Point	16	39	42	97
Revised	43	32	20	95
Concise	12	44	38	94
Descriptive	29	21	27	77
Long	10	30	30	70
Problem statement	17	25	23	65
Real	22	16	22	60
Thematic	16	25	13	54
Vague	25	13	16	54
Short	9	19	22	50
Total	1307	1449	1389	4145

Table 4 Chi-squared results on word frequency

Term	Chi-squared *p* value
Need	0.53
Clear	0.73
Insight	0.38
Stakeholder	0.43
Specific	0.024
User	0.0005
Solution	0.057
Focus	0.17
Action	0.46
Detail	0.02
Format	0.18
Define	0.53
Point	0.005
Revised	0.006
Concise	0.0004
Descriptive	0.32
Long	0.008
Problem Statement	0.64
Real	0.40
Thematic	0.18
Vague	0.05
Short	0.10

investigate potentially significant differences in word counts between groups, we noted limited significant differences between groups (see Table 4).

Despite the lack of statistically significant differences between groups, our analysis of the word counts did reveal an interesting trend between groups; namely, the mixed-quality group had more mentions of the identified words than the other two groups (i.e., the mixed-quality group had 1449 mentions of the identified words, the high-quality group had 1307, and the low-quality group had 1389). Additionally, the mixed-quality group showed a greater range in the counts for words coded (e.g., word counts in the 100 s, 90 s, 80 s…10 s) while the high- and low-quality groups were more distinct (e.g., word counts in the 100 s and then fewer in the 50 s, 30 s, and 20 s).

The second analysis of student comments consisted of categorizing the LbE rationales provided by students from each group as either positive, negative, or neutral. Our findings showed that, in all groups, there were more than twice as many positive comments (High-Quality Group = 483, Mixed-Quality Group = 443, Low-Quality Group = 495) as neutral comments (High-Quality Group = 165, Mixed-Quality Group = 227, Low-Quality Group = 197) and even fewer negative comments (High-Quality Group = 95, Mixed-Quality Group = 140, Low-Quality Group = 152) (See Table 5). Further analysis showed the group only exposed to high quality examples (High-Quality Group) had significantly fewer negative comments than the other two groups—an intuitive finding given the high-quality nature of the items they compared. Overall, the counts among the items suggest that

Table 5 Comment sentiment by group

	Positive	Negative	Neutral
High-quality group	483	95	165
Mixed-quality group	495	152	197
Low-quality group	443	140	207

students were more likely to justify their judgement decisions with positive comments than critical ones.

The third qualitative analysis of student LbE rationale comments compared the comments provided by students during LbE to the feedback that they received on their own project from their peers (see Table 3). Specifically, all comments were coded as positive, negative, or neutral and the counts of comments were analyzed for any potential differences. Again, our analysis demonstrated no significant difference between groups. All groups received four times as many positive comments as negative comments during the POV evaluation—a finding which was also matched in the comments provided by these students during the LbE intervention at the beginning of the project. While our analyses revealed very few differences between groups, there were several findings of interest that hinted at how students engaged in the LbE process. For example, students' LbE comments generally followed a theme in which their feedback centered on one specific aspect of a POV across all examples evaluated through LbE. For example, feedback provided by one student included the following:

"The other does not directly mention **groups** involved in the [POV] statement.",
"The other does not identify any **user groups** or **stakeholders**.",
"The other does not specify the **user group**.",
"There are no **user groups** specified in the other.",
"The other does not specify **user groups**." and
"This one is more specific in its plan and its **stakeholders**."

In each instance, this student evaluated POV statements during the LbE intervention and focused their feedback solely around user groups. While user groups are an integral part of the overall POV statement creation, they are just that—one part. This theme of feedback revolving around one aspect/idea was common across many of the evaluations made by students.

Another trend that we found interesting was a level of quality conditioning that appeared to impact student judgments. For example, we found that students approached the judgment process relatively, meaning, they made judgements based on the caliber of examples they were exposed to as opposed to a larger view of potential quality. Specifically, students who were only exposed to high quality examples sometimes concluded that "Both of these are poor," and students who were only exposed to low quality examples concluding that "Both of these were very good".

Discussion and conclusion

While previous analyses clearly support the use of LbE in the classroom, there was no understanding of what quality of items should be included in LbE. Our attempt to understand the types of examples that should be presented to students provides potential direction

for educational research and practice; while our analysis showed no statistically significant differences in student achievement for students shown only high-quality, low-quality, or mixed-quality items, our qualitative investigation of student comments did reveal interesting differences in student experience based on group. Fully recognizing limitations in our study (e.g., because of previous findings suggesting the positive impact of LbE we determined to engage all students in some level of LbE experience), we nevertheless believe there is significance and utility in several different findings from this study. First, we found it counter-intuitive that differences in student performance on the POV assignment were not significant despite being exposed to differing quality of example POVs during LbE. We hypothesized that the students shown only high-quality examples would "rise to the occasion" and outperform their peers; or, that students shown mixed-quality examples would better differentiate the nuances of quality and apply that to their own learning. The quantitative findings did not confirm these hypotheses and we wonder if perhaps the influence of LbE on student learning was isolated from item quality? It is also possible that the differences in quality of items shown to students (e.g., high vs. low vs. mixed) were not distinct enough to provoke a difference in student capacity and transfer. Or, we wonder if some of the students (e.g., those in the high-quality or low-quality only groups) may not have been able to fully appreciate "excellence" and the range of quality, when shown only a limited variety of item (e.g., Kimbell, 2018 discusses the value in exposure to varying levels of quality). Additional investigation into why, and how, the difference in student experience translates into future work is needed to better understand the nuances of LbE.

The first step of our qualitative analysis into the experiences of students in each of the three groups supported our quantitative findings as there were few significant differences in the word counts from each of the groups. The relative consistency among word counts between groups suggests that students were perhaps able to focus on the assignment specifications despite the differences in quality of the POVs displayed. Further, the similarities support the idea that the benefits of the LbE experience may be independent of quality of item shown. However, as we noted, although not statistically significant, there were differences in the word counts between groups. Specifically, the mixed-quality group showed a higher word count and a greater range in counts for the identified words. These differences may be indicative of their exposure to a wide range of quality in items and a subsequent greater ability to recognize the nuances of quality in items (e.g., a more "rounded" view of quality in items displayed).

Our second qualitative analysis centered on the overall sentiment of student's comments (positive, negative, or neutral). It was not surprising to us that most student comments were positively coded as our classroom experience suggests students are more comfortable commenting on positive aspects during critique than they are negative. Importantly, those students who only view high-quality items had significantly fewer negative comments than their peers from the other groups—an intuitive finding given the items they viewed. However, we also note that this may be an important consideration for teachers; if students are more comfortable providing positive comments, perhaps teachers should consider only showing high-quality items. Conversely, perhaps there is distinct value in identifying and vocalizing areas for improvement—something to consider when identifying the items for student comparison in the classroom.

Our third qualitative analysis revealed other interesting trends—while we noted no significant relationship between the comments provided by students and those received later, we did note a pattern in the feedback provided by students. Specifically, we recognized that students appeared to become fixated on one element of the larger assignment (e.g., the user) and used this element for all their feedback. This raises a host of questions around design fixation,

transfer, and creativity; for example, if students are providing all their feedback around a single element of the larger assignment are they truly identifying which item is "better" or are they being blinded by the element of choice? Why are students fixating on distinct part of the assignment—is it a matter of laziness, understanding, or comfort? Does this fixation on one aspect of the larger assignment translate into a better, or worse, ability later on? Additional investigation into this finding, and the ramifications for LbE and design education, is needed to better understand what is happening and why.

Finally, we also noted in our findings that students experienced a level of what we termed "quality conditioning." Specifically, students exposed to only high-quality items still deemed sets of items as both being "bad" while those exposed to only low-quality items commented that "both of these items were good." We are not sure how or why this happened—perhaps this relates to a fixation on one element of the overall design or, alternatively, perhaps they became so conditioned by the items they viewed that their perception of quality was changed. Alternatively, perhaps this was a function of being exposed to a limited variety of items (e.g., only high-quality items). Additional research is also needed in this vein to investigate this phenomenon.

Overall, our analyses suggests that while the types of examples presented to students in the process of LbE does not have a statistically significant influence on their future performance on a similar assignment, there were—sometimes subtle—differences in their experience. Additional exploration into the identified differences, and similarities, is needed to further expand our understanding of the experiences of students in these settings. Further, this study provides additional insight into aspects of students' learning using LbE and, in addition to questions raised around how students' initial LbE experience affects their academic achievement as well as their expectations, we also uncovered other questions about how and why students engage with the examples in different ways.

We anticipate that interviewing students and teachers about their experiences using LbE could provide greater insight into how teachers facilitate learning, and the thought processes students participate in while engaged in LbE. Additional work with varied quality in examples shown would be useful—especially when combined with qualitative exploration of student experience, thinking, and decision-making. Exploratory research that identifies different types of learning activities would also allow further research into optimal implementations of LbE.

Funding This material is based upon work supported by the National Science Foundation under Grant 2101235.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Disclaimer Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Baker, S. E., & Edwards, R. (2012). How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. *National Centre for Research Methods*, pp. 1–42.

Baniya, S., Chesley, A., Mentzer, N., Bartholomew, S., Moon, C., & Sherman, D. (2019). Using adaptive comparative judgment in writing assessment: An investigation of reliability among interdisciplinary evaluators. *Journal of Technology Studies, 45*(2), 24–45.

Bartholomew, S. R., Strimel, G. J., Garcia Bravo, E., Zhang, L., & Yoshikawa, E. (2018). Formative feedback for improved student performance through adaptive comparative judgment. *Paper presented at the 125th ASEE conference*, Salt Lake City, Utah.

Bartholomew, S. R., & Yoshikawa, E. (2018). A systematic review of research around adaptive comparative judgment (ACJ) in K-16 education. *2018 CTETE Monograph Series*. https://doi.org/10.21061/ctete-rms.v1.c.1.

Bartholomew, S. R., & Yauney, J. (2022). The impact of differentiated stimulus materials in learning by evaluating. *Pupils' Attitudes Towards Technology 39th Annual Conference*, St. John's, Canada, 2022. https://par.nsf.gov/servlets/purl/10340851.

Bartholomew, S. R. (2017). Assessing open-ended design problems. *Technology and Engineering Education Teacher, 76*(6), 13–17.

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*. https://doi.org/10.1007/s10798-020-09639-1

Bartholomew, S. R., Strimel, G. S., & Yoshikawa, E. (2018b). Using adaptive comparative judgment for student formative feedback and learning during a middle school open- ended design challenge. *International Journal of Technology and Design Education, 29*(2), 363–385.

Bartholomew, S. R., 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.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment* (p. 36). Cambridge Assessment.

Buckley, J., Seery, N., & Kimbell, R. (2022). A review of the valid methodological use of adaptive comparative judgment in technology education research. *Frontiers in Education, 7*, 787926. https://doi.org/10.3389/feduc.2022.787926

Caniglia, J. (2020). Promoting self-reflection over re-teaching: Addressing students misconceptions with "my favorite no." *Journal of Mathematics Education, 5*(2), 70–78.

Canty, D. (2012). *The impact of holistic assessment using adaptive comparative judgment of student learning*, PhD Thesis, University of Limerick, Ireland.

Collins, A. (2022). *Cognitive apprenticeship*. Retrieved April 11, 2022 from https://www.isls.org/research-topics/cognitive-apprenticeship/.

Collins, R. (2014). Skills for the 21st Century: Teaching higher-order thinking. *Curriculum and Leadership Journal, 12*(14), 1–8.

Dam, R. F., & Siang, T. Y. (2020). *Stage 2 in the design thinking process: Define the problem and interpret the results*. The Interaction Design Foundation. Retrieved November 15, 2021, from https://www.interaction-design.org/literature/article/stage-2-in-the-design-thinking-process-define-the-problem-and-interpret-the-results.

Johnson, C. C., Sondergeld, T. A., & Walton, J. B. (2019). A study of the implementation of formative assessment in three large urban districts. *American Educational Research Journal, 56*(6), 2408–2438. https://doi.org/10.3102/0002831219842347

Johnston, O., Wildy, H., & Shand, J. (2019). A Decade of teacher expectations research 2008–2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education, 63*(1), 44–73. https://doi.org/10.1177/0004944118824420

Kimbell, R. (2018). Constructs of quality and the power of holism in *Pupils attitudes towards technology 36th Conference Proceedings,* pp. 181–186.

Kimbell, R. (2012). The origins and underpinning principles of e-scape. *International Journal of Technology and Design Education, 22*, 123–124.

Kimbell, R. (2021). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-021-09654-w

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*, 772832.

Miksza, P. (2011). A review of research on practicing: Summary and synthesis of the extant research with implications for a new theoretical orientation. *Bulletin of the Council for Research in Music Education, 190*, 51–92. https://doi.org/10.5406/bulcouresmusedu.190.0051

Pollitt, A. (2004). *Let's stop marking exams.* Retrieved from http://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf.

Pollitt, A. (2015). On 'reliability' bias in ACJ. *Cambridge Exam Research.* Retrieved April 23, 2020 from https://www.researchgate.net/publication/283318012_On_'Reliability'_bias_in_ACJ.
Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice, 19*(3), 281–300.
Rangel-Smith, C., & Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment. *In 36th pupils' attitudes towards technology conference*, Athlone, Ireland, pp. 378–387.
Robertson, S., Humphrey, S., & Steele, J. (2019). Using technology tools for formative assessments. *The Journal of Educators Online*. https://doi.org/10.9743/jeo.2019.16.2.11
Saldaña, J. (2015). *The coding manual for qualitative researchers*. SAGE.
Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*(4), 759.
Seery, N., & Canty, D. (2017). Assessment and learning: The proximal and distal effects of comparative judgment. *Handbook of Technology Education*. https://doi.org/10.1007/978-3-319-38889-2_54-1
Sherman, D., Mentzer, N., Bartholomew, S., et al. (2022). Across the disciplines: our gained knowledge in assessing a firstyear integrated experience. *Int J Technol Des Educ*, *32*, 1369–1391. https://doi.org/10.1007/s10798-020-09650-6.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.
Wible, S. (2020). Using design thinking to teach creative problem solving in writing courses. *College Composition and Communication, 71*(3), 399–425.