Investigating the Impacts of Differentiated Stimulus Materials in a Learning by Evaluating Activity

108th Mississippi Valley Technology Teacher Education Conference, Nashville, TN
November 18-19, 2021

Scott Bartholomew[1], Jessica Yauney[1], Nathan Mentzer[2], Scott Thorne[2]
[1] Brigham Young University
[2] Purdue University

Abstract
Previous work has demonstrated the capacity for significantly influencing student achievements through engagement in learning by evaluating (LbE) activities via adaptive comparative judgment (ACJ). Expanding on this research, which highlighted a significant difference between students who engaged in learning by evaluating and those who did not, this effort specifically investigated the impact of differentiating the quality of examples students engaged with while completing LbE exercises. In this research, university design students (N = 468) were assigned to one of three treatment groups; each group completed LbE activities with examples of differing quality within their LbE comparisons prior to designing. Specifically, some students only evaluated high quality examples, others only evaluated low quality examples, and the third group of students evaluated a mix of high- and low-quality examples. Following the LbE activities, students completed the assigned work and student achievement on the project was subsequently analyzed to determine if there was a difference between groups. No significant difference was found between the groups in terms of achievement. Additional analysis was completed on student evaluation rationales to explore differences in student behavior based on intervention grouping.

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

Introduction

While K-12 classrooms engage with learning and assessing daily, assessment is often viewed as the role of the teacher and only having limited effect on student learning through teaching choice to revisit or expand on earlier material (Johnson et al. 2019). While assessment practices have improved over time (e.g., greater access to technology for facilitating assessment) (Robertson et al. 2019), relatively little has changed in terms of students' participation in assessment processes - for the most part, students submit work and teachers evaluate this work and then assign grades. This formulaic approach to assessment often coincides with assessment signaling *the end* of student learning as opposed to one more step in the larger learning journey (Bartholomew et al., 2020). Innovative work with assessment has demonstrated the potential of assessment to play a much larger role in students' learning as they are invited to engage in the assessment process and subsequently revisit/review/revise their own work. Research into 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 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 (Bartholomew, Mentzer, Jones, & Sherman, 2020).

Adaptive Comparative Judgement. Adaptive comparative judgment (ACJ) is a method of assessing items (e.g., student work) by making a series of comparative evaluations. In ACJ an individual views pairs of items and determines, based on an identified criteria, which is better. This process is repeated iteratively with different pairing of items and the outcome is a rank-ordered list of all included items based on pre-selected criteria and determined by the evaluative comparative judgments. ACJ derives from comparative judgment (CJ) which was developed by psychometrician Thurstone (1927) and later refined by academic Alastair Pollitt (2004, 2012). In CJ, Thurstone and Pollitt proposed an alternative approach to assessment through rubrics or other criterion-based approaches using a series of comparative judgments which could be completed to qualitatively sort a body of educational items. This idea 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). Both CJ and ACJ methods use a holistic approach to completing comparative evaluations instead of a more-commonly used rubric-based evaluation consisting of a series of sub-criteria and point-values. In ACJ the comparative process is theoretically shortened and refined through an underlying algorithm which serves to sort and pair items *adaptively* to produce optimized results (in terms of both reliability and efficiency). Other data outside the final rank ordering of items may also be collected during this process including judgment times and judge's rational/comments (Bartholomew 2017; Pollitt 2004, 2012). Previous research has shown high reliability levels (Baniya et al. 2019, Bramley 2015), a simpler assessment process (Kimbell, 2021), and greater ease of integrating assessment feedback from multiple assessors (Bartholomew and Yoshikawa 2018; Kimbell 2012b) 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).

Learning by Evaluating. For the first decade following its proposal (Pollitt 2004), ACJ was primarily used as an alternative approach aimed at improving assessment reliability (see Bartholomew & Jones, 2020). However, researchers (e.g., those in the UK (Kimbell), Ireland (Canty), and the USA (Bartholomew)) recognized the potential for utilizing the ACJ assessment approach as a learning tool for students. Specifically, Bartholomew, Mentzer, Jones, & Sherman (2020) coined the term *Learning by Evaluating* (LbE) to describe a process wherein students use ACJ to view and evaluate examples of work *prior* to engaging in an assignment themselves. As a learning intervention, LbE is a specific application of ACJ where students engage as judges in ACJ with the primary goal of learning through the evaluation of previously-submitted artifacts. 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, Mentzer, Jones, & Sherman, 2020; Bartholomew and Strimel 2019; Bartholomew et al. 2018b; Bartholomew et al. 2018a; Seery and 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 (see Bartholomew & Jones, 2020).

Research Question. As recent research has shown the value of using this assessment tool as a learning experience, questions around the potential to modify or enhance this LbE experience have risen (Bartholomew et al. 2020). Specifically, Bartholomew et al. (2020) highlighted the potential for improving student learning by intentionally influencing, and varying, the quality of work presented to students during the LbE comparison – if LbE has demonstrated a significant impact on student learning through the use of randomly-assigned evaluations, can this learning 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: What is the impact, if any, on student learning of differentiated stimulus materials in a learning by evaluating activity?

Based on previous work and findings (Bartholomew et al. 2020), we hypothesized that engaging students in LbE with only high-quality examples would set a high standard and lead to better educational outcomes - a hypothesis we deemed appropriate based on our observations, 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 would lead to lower outcomes following the LbE; finally, we hypothesized that engaging students with mixed quality examples may lead to mediocre educational outcomes. Relatedly, we further hypothesized that engaging students with LbE using mixed-quality examples could potentially lead to high educational outcomes—potentially even higher than those obtained by LbE with only high-quality comparisons—based on research (Miksza 2011) that shows that the act of comparing/contrasting cases engages higher-order thinking and may lead to increased educational attainment.

Methods

To better understand the nuances of item inclusion in LbE scenarios, we built on previous research conducted by Bartholomew, Mentzer, Jones, & Sherman (2020) with university design students. Specifically, we engaged 468 students in LbE as part of a larger design project contained within a required course on design thinking. These students worked in groups (N = 112) to complete an 8-week design project as part of this larger course and, as part of their 8-week design project, students learned how to create Point of View (POV) Statements as a means of framing their designing. POV statements are a design tool with three components: a user, a need, and a unique insight (Wible, 2020) which provide both the impetus for, and the boundaries within, which the larger design experience happens (Dam & Siang, 2020). Similar to Bartholomew, Mentzer, Jones, & Sherman (2020), we engaged students in LbE of POV statements before they worked in groups to create their own statement. Specific to our research question around what influence, if any, differing quality items included in LbE may have on student learning, all student groups were assigned to one of three treatment conditions. We intentionally opted to provide all students with the LbE experience based on previous research which showed that LbE provided better educational outcomes for students; therefore, this research did not include a traditional "control group" with no LbE experience in class. The student grouping was as follows:

- Group A: Students who only evaluated high quality examples
- Group B: Students who only evaluated low quality examples
- Group C: Students who evaluated a mix of both low- and high-quality examples.

During the LbE intervention process, which took approximately 20-minutes of class time, all students were asked to complete six comparisons (looking at up to 12 different examples of previously submitted student group POV statements) where they selected which of two displayed POV statements was better and provided a short explanation for their decision (see Figure 1).

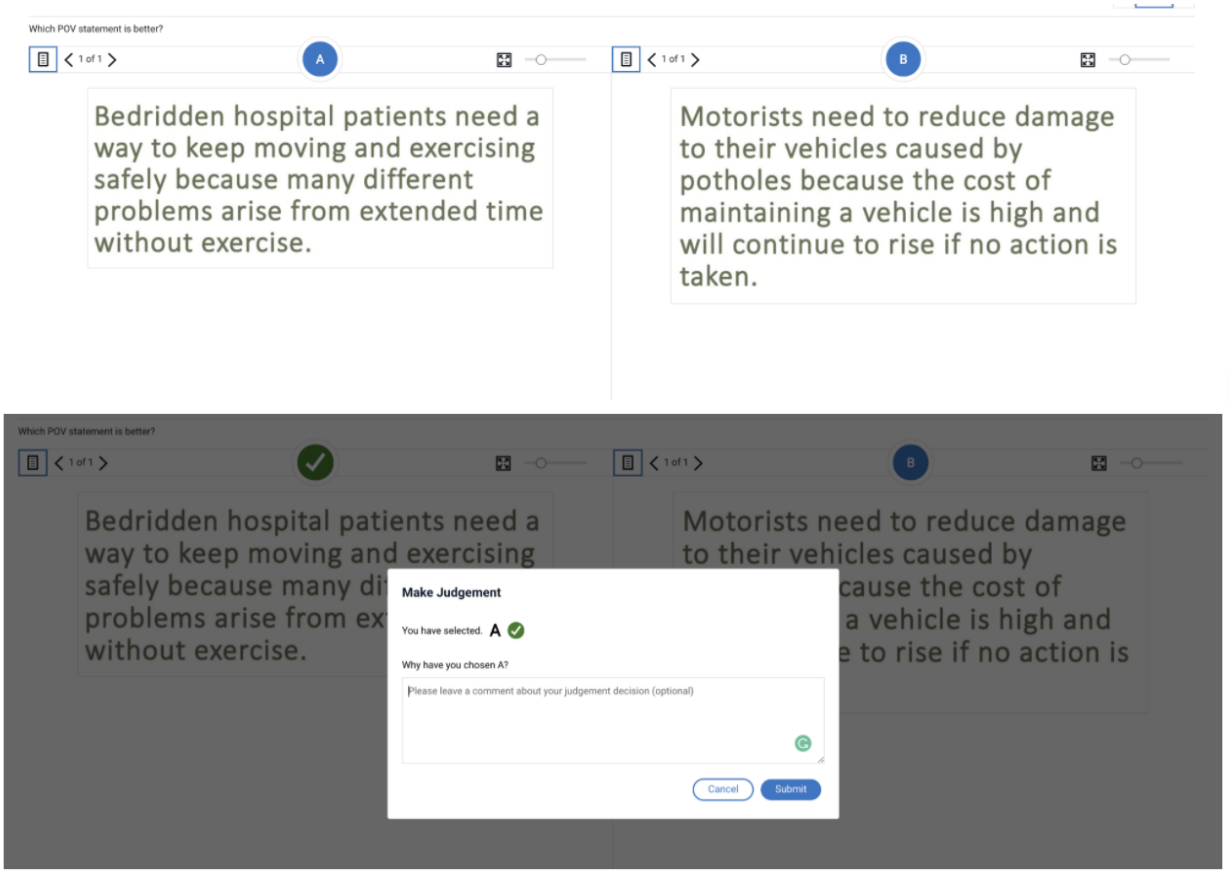

Figure 1 – *LbE POV student view*

Outside the 20-minute LbE intervention, all other classroom procedures, routines, schedules, and assignments were identical to previous semesters. Further, the students' LbE intervention assignment (Group A, B, or C) happened at the group level (as opposed to a class or student level) as a means of addressing potential differences existing between classes, sections, or course instructors; in this way, all students in a POV group received the same LbE intervention (level of quality work viewed during the LbE intervention) but each POV group within a class was independently, and randomly, assigned to the treatment condition. Students, and course instructors, were unaware of the treatment condition assignments (A, B, or C) during the LbE intervention as these were only visible to the researchers following the intervention.

The feedback provided by students—framed as a rationale or short explanation surrounding their decision to choose one item over another—was collected through the ACJ software interface (RMCompare, compare.rm.com) and sorted for later analysis. Following the LbE intervention, during which students viewed work from a previous semester of differing quality based on their randomly assigned grouping, all students worked in groups to create their own POV statements for a problem of their choice. This POV creation was part of a larger design experience (encompassing ~8 weeks of class time) wherein students experienced a design cycle and worked to provide a solution for an identified problem. Following the design experience, all students submitted their POV statements for assessment

The POV statements (N=112), created by students from each of the POV groups, were then added to a separate ACJ session for evaluation. All 468 students were enrolled as judges again to evaluate these newly created POVs in this second LbE session. These POV statements were evaluated by the students using the ACJ software and the resulting rank order was used as a learning indicator for each of the students in line with our stated research question and hypotheses.

Our decision to utilize the students as evaluators during this second ACJ session was intentional - previous research has consistently demonstrated high levels of reliability and correlation between student judges and professional judges/instructors (Bartholomew & Jones, 2020). Further, in this instance, students were evaluating POV quality following an entire semester course on the given topic and given the large number of items (112), the burden of completing a significantly high round of judgments was alleviated by utilizing all enrolled students (N=468) as judges. The resulting rank order from these students, obtained following 39 rounds of judgment, was high (r = .83) thus lending credence to our approach and the consistency of judgment decisions made by students.

This initial data collection included the gathering of all student judgments, the final rank order of POV statements based on the evaluation of students, and all students' comments provided while evaluating the pairs of POVs. Once completed, statistical analyses were completed on the quantitative data through statistical analysis software (SPSS) 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.

Next, additional qualitative analysis consisted of three exploratory analyses completed on students' comments collected during their LbE experience. In the first analysis, we chose to analyze the prevalence of terms provided by students in the LbE comments to potentially identify trends and better understand what students may be learning in the process of LbE. This analysis aligned with our general research inquiry around the nuances of the LbE experience and followed recommendations by Saldaña (2015) for attribute coding with frequency counts with the intent of further exploring our research question and the associated hypotheses. 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 was sorted in line with the intervention groups (A, B, or C) 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 illuminating and triangulating findings derived from the other analyses performed during this effort.

Our second qualitative analysis also followed Saldaña's (2015) recommendations – this time in line with suggestions for attribute coding using thematic categories. In line with our hypotheses, and 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. For each of the 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 comments that included both positive and negative comments were coded as neutral.

The third qualitative analysis sought to understand the relationship, if any, between the student rationales provided during the LbE intervention at the beginning of the intervention, and the comments provided by other students on the POV statements created during class. Specifically, we sought to identify if students' comments provided during the LbE experience related to the feedback they received on their own projects at the conclusion of the POV creation process. In this third analysis, 20 students were randomly selected from each treatment (Group A, B, and C) and an analysis of the rationales provided during the LbE experience was completed in line with recommendations from Baker & Edwards (2012) for qualitative analysis. For each of these students the rationales given by peers during the evaluation of their POVs was also gathered. 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. Additional data conditioning was performed to ensure any feedback included in this analysis was specific to the identified student's POV. Finally, each of these comments was qualitatively analyzed to explore the potential for patterns between student LbE rationale and the rationale provided by peers during the final POV.

Findings

Based on previous findings which have shown that students who use LbE have better academic outcomes than those who do not use LbE, we determined to investigate the potential, if any, to influence student learning outcomes by intentionally differing the quality of items evaluated during LbE. Our analysis revealed no statistically significant difference between students who were exposed to high, low, or mixed quality examples (p = .809). Specifically, the difference between high quality (M = .13, SD = .836), low quality (M = -.07, SD = .970) and mixed quality (M = -.04, SD = 1.12) groups was not significant either overall or between each of the Groups (A, B, or C). 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 and the associated hypotheses, we next investigated the potential for significant differences in the qualitative data provided by students' comments on the six example POVs 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 qualitative analysis consisted of analyzing the words included in the student rationale provided during the LbE intervention. This was done in an effort to further investigate our research question around what impact, if any, the differences in item quality viewed during LbE may have on students learning we investigated the frequency with which these different terms appeared in the LbE rationales for students within each of the groups (A, B, or C). General overall investigation 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. Following 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. We noted that within these groups of commonly appearing words there was no significant difference between groups in terms of counts. This comparable use of content specific vocabulary and writing critique between groups further supports our quantitative finding of no significant difference in the outcomes of student groups in terms of learning.

The second qualitative analysis 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 (Group A = 483, Group B = 443, Group C = 495) as neutral comments (Group A = 165, Group B = 227, Group C = 197) and even fewer negative comments (Group A = 95, Group B = 140, Group C = 152). Further analysis showed that the only significant difference amongst groups was that the group only exposed to high quality examples (Group A) 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 students were more likely to justify their judgement 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. 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 seeing with even students who were only exposed to high quality examples sometimes concluding 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

While previous analyses clearly support the use of LbE in the classroom, this attempt to better understand the types of examples that should be presented to students provides much less clear direction for education practice. Our analysis suggests that students may experience positive learning impacts regardless of the types of examples given and provides some potential explanations for this phenomenon. Alternatively, since there was no control group in our study, we cannot conclusively argue that LbE is an effective learning intervention regardless of the quality of items included in the comparisons. Further, there are many other questions that still exist surrounding how to most effectively use LbE. Given that there was no statistically significant effect on learning from the quality of the examples given, researchers should continue to analyze other aspects of LbE such as the effect of prompts, teacher introductions, and peer review on student learning.

The fact that students exposed to only high-quality examples made negative comments and that those exposed to only low-quality examples made positive comments raises questions about how students' expectations are shaped by what they see. A potential avenue for future research includes exposing students to new examples after the initial evaluations to see how students' opinions differ based on the examples they are exposed to.

Conclusion

Researchers extended earlier research that shows the value of LbE to improve student learning. Students engaged with LbE with only high-quality examples, only low-quality examples, or a mix of both high- and low-quality examples. No differences were found in overall student outcomes based on the quality of examples they evaluated. However, data does suggest that students' expectations were influenced by the other examples they viewed. Data also showed that student comments most often focused specifically on select elements of the POV statements that students were previously taught.

Overall, our analysis suggests that the types of examples presented to students in the process of LbE does not have a significant impact on student learning. However, this study also provides additional insight into aspects of students' learning using LbE. While questions were raised regarding how students' initial LbE experience affects their academic achievement as well as their expectations, we uncovered other questions about how and why students engage with the examples in different ways.

In order to improve understanding of LbE and to better facilitate learning activities, there are many areas for future research. Interviewing students and teachers about their experiences using LbE could provide greater insight into how teachers facilitate learning and what thought processes students participate in. Exploratory research that identifies different types of learning activities would allow further research into optimal implementations of LbE.

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

Bartholomew, S. R. (2019). Teacher led assessment of learning to student led assessment as learning. *Beyond multiple choice conference*, Washington DC.

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.erl.lib.byu.edu/10.1007/s10798-020-09639-1.

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., 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., & 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., 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. *Cambridge Assessment*, Cambridge, 36.

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. (2007). E-assessment in project e-scape. *Design & Technology Education: An International Journal*, 12(2), 66–76.

Kimbell, R. (2012a). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, 22(2), 135–155.

Kimbell, R. (2012b). 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

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/image s/109719-let-s-stop-marking-exams.pdf.
Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281–300.
Pollitt, A. (2015). On 'Reliability' bias in ACJ. *Cambridge Exam Research.* Retrieved 23, April 2020 from https://www.researchgate.net/publication/283318012_On_'Reliability'_bias_in_ACJ.
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*, *16*. https://doi.org/10.9743/jeo.2019.16.2.11
Saldaña, J. (2015). *The coding manual for qualitative Researchers*. Thousand Oaks, CA: SAGE.
Seery, N., & Canty, D (2017). Assessment and learning: The proximal and distal effects of comparative judgment. *Handbook of Technology Education*, Springer. DOI: 10.1007/978-3-319-38889-2_54-1.
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. Retrieved from http://erl.lib.byu.edu/login/?url=https://www-proquest-com.erl.lib.byu.edu/scholarly-journals/using-design-thinking-teach-creative-problem/docview/2378096479/se-2?accountid=4488