Examining the Potential of Adaptive Comparative Judgment for Elementary STEM Design Assessment

By Scott R. Bartholomew, Greg J. Strimel, Liwei Zhang, and Jessica Homan

ABSTRACT

STEM education practices and approaches have been emphasized in recent years at the elementary school level. The emphasis on STEM integration at the elementary level has stressed learning, motivation, and 21st-century skills as positive outcomes. Despite this emphasis, elementary level teacher assessment practices for open-ended STEM design challenges are not clearly established. Additionally, little is known about the teacher workload associated with various forms of assessment connected with these activities. Therefore, the researchers collected and examined data from four teachers and 100 elementary school students engaged in three STEM design problems. Teachers assessed student work using traditional approaches and a relatively new approach called adaptive comparative judgment (ACJ). The time teachers spent assessing student work using the two forms of assessment, the scores received through traditional assessment approaches, and the rank order of student work from the ACJ assessment were collected. The data analysis revealed key similarities and differences, in the time required for assessment and the outcome of traditional and ACJ assessment approaches.

Keywords: adaptive comparative judgment, elementary school STEM, design, assessment

INTRODUCTION

Efforts aimed at increasing science, technology, engineering, and mathematics (STEM) education at the elementary (grades K-6, ages 5-12) level have spread dramatically in recent years (Daugherty, Carter, & Swagerty, 2014; Dejarnette, 2012; Murphy, 2011). Legislation, standards, curriculum, professional development, funding, and a variety of other resources have all been employed towards this end (Daugherty et al., 2014). Specifically, the development of the *Next Generation Science Standards* (NGSS Lead States, 2013), along with several initiatives geared toward STEM participation and younger students (i.e., *Engineering is Elementary, Engineering by Design, Project Lead the Way Launch, Teach Engineering, STEM: It's Elementary*) have all led to an increased emphasis on STEM integration in elementary school classrooms. As elementary school teachers and administrators work to integrate STEM into their classroom, integration may often take shape in the form of problem-based, project-based, and design-based learning activities (Laboy-Rush, 2011), which engage students in hands-on activities with designing, prototyping, building, testing, and evaluating solutions to a posed problem. Advocates for implementing STEM activities, and the corresponding pedagogical approaches, posit that this integration will improve student motivation, achievement, and help students develop necessary 21st-century skills and competencies for success in life (Daugherty et al., 2014; Laboy-Rush, 2011; National Academy of Engineering [NAE] & National Research Council [NRC], 2014; NGSS Lead States, 2013).

However, in tandem with these curricular changes and emphases, teachers have often grappled with questions around the appropriate approaches to assess these open-ended activities (Bartholomew, 2017; Bartholomew, Nadelson, Goodridge, & Reeve, 2017; Kimbell, 2007, 2012; Pollitt, 2012). This difficulty in assessment is not confined to elementary school classrooms; the open-ended nature of STEM problem-, project-, and design-based learning activities implies there is no single, correct answer for teachers to use in assessment (Bartholomew, 2017; Bartholomew & Strimel, 2017; Pollitt, 2012). In conjunction with this difficulty, a variety of assessment approaches have been developed that employ various tools and techniques to try and improve these assessment situations; these include approaches such as rubrics, technology assessment platforms, and/or questionnaires (Bartholomew, 2017; Denson, Buelin, Lammi, & D'Amico, 2015; Diefes-Dux, Moore, Zawojewski, Imbrie, & Follman, 2004; Pollitt, 2004).

One specific method for assessing open-ended problems, an approached called adaptive comparative judgment (ACJ), has proven especially reliable, valid, and effective with these open-ended problems at the middle school, high school, and higher education levels (Bartholomew, 2017; Bartholomew & Yoshikawa, 2018; Pollitt, 2012; Seery, Canty, & Phelan, 2012; Tarricone & Newhouse, 2016). ACJ, as an approach to assessment, has been embodied in several web-based technology tools (e.g., *CompareAssess*), but it has not been previously tested with elementary school teachers in the United States (Bartholomew & Yoshikawa, 2018). Further, ACJ—an approach developed with the express intent of a group of graders/judges looking at student work (Pollitt, 2004)—has not been tested with individual assessors completing the entire process (Bartholomew & Yoshikawa, 2018). While we fully recognize that ACJ, as an approach, and *CompareAssess*, as a tool, were not intended to be used by individual judges, we also recognize that the educational paradigm and practices existing in many countries are centered on individual teachers assessing student work without the input of other assessors or judges (DigitalAssess, 2017).

Considering the increased emphasis on STEM education at the elementary school level, the challenges associated with the assessment of student work, the lack of research with ACJ at the elementary school level, and the uncertainty about the potential of an individual teacher using ACJ for assessment, this research sought to specifically investigate the teacher workload of individual elementary teachers while conducting assessments using both a traditional rubric approach and ACJ. This study specifically examined teacher experiences with the assessment of student work from 100 elementary school students (50 kindergarten [ages 5-7] and 50 fourth-grade [ages 8-10] students) using traditional rubric-centered approaches to assessment as well as ACJ. The data collected includes the score received through traditional rubric-based assessment and the rank order derived through ACJ assessment. Additionally, the time needed for teachers to complete each form of assessment for the student work and the teacher responses to a post-study questionnaire were both solicited to explore these implications and ideas.

PURPOSE OF THE STUDY

While the emphasis on STEM education practices, approaches, and techniques at the elementary school level has increased over recent years (Daugherty et al, 2014; Nadelson et al., 2013), less has been done to investigate different approaches to assessing open-ended design problems. ACJ, a new approach with high levels of reliability, appears to be a suitable approach for elementary school STEM settings, but little research has been done to test this potential and the associated classroom implications. Therefore, the purpose of this research was to examine the potential of using ACJ for assessing elementary school STEM design activities with individual teachers acting as judges.

RESEARCH QUESTIONS

Three specific research questions, which were used to guide this study and explore the overall potential and opportunities around ACJ in elementary school STEM assessment, were established:

RQ_1: What relationship exists, if any, between student achievement measures obtained through ACJ and rubric-based assessment approaches for elementary school STEM design activities?

RQ_2: What are the implications for teacher workload associated with ACJ and rubric-based assessment approaches to elementary school STEM design activities?

RQ_3: What are the practical implications of implementing ACJ assessment for elementary school STEM design activities by individual teachers?

Elementary STEM Design Education

The importance of STEM education, as early as elementary school, has been highlighted increasingly in recent years (Archer et al., 2012; Daugherty et al., 2014; DeJarnette, 2012; Kuenzi, 2008; Murphy, 2011). Kuenzi (2008) noted that STEM education, and the achievement of students in STEM subject areas, is critical in light of ensuring continued scientific and technological developments in years to come. Leading standards for STEM areas such as the *Standards for Technological*

Literacy (ITEA/ITEEA, 2000/2004/2007) and the *Next Generation Science Standards* (NGSS, 2013) emphasize STEM concepts, principles, and literacy at the elementary level. DeJarnette (2012) posited that early exposure to STEM at the elementary level might lead to an increase in students' interest in STEM fields later in life. Furthermore, advancing STEM education may also provide students with opportunities to develop 21st-century skills, such as communication, problem solving, and systems thinking, while also increasing their understanding of important issues such health, energy efficiency, and environmental quality (Bybee, 2010).

STEM Design Activities

Classroom activities that incorporate STEM principles and concepts revolve around problem- and project-based learning scenarios where students often work in groups to design a solution to a problem (Laboy-Rush, 2011). Relatedly, the *Standards for Technological Literacy* suggest that elementary STEM education should provide students with diverse opportunities to address their natural curiosity and inventive thinking skills, and to develop their skills in designing, planning, making, and presenting solutions to technological problems (ITEEA, 2000/2004/2007). With the goal of STEM literacy, elementary teachers and educators can integrate problem-based learning opportunities which involve students in hands-on events where they design, make prototypes, test, evaluate, and document their solutions to real-life problems (Reeve, 2015). These types of learning situations have been proven effective in developing critical thinking, promoting student interest, and increasing opportunities for interactivity and innovation (DeJarnette, 2012).

Elementary STEM Design Assessment

Although there is an increased emphasis on STEM teacher training, knowledge of content, and pedagogy (Daugherty et al., 2014), the assessment of these open-ended STEM activities continues to be challenging for teachers (Bartholomew, 2017; Bartholomew & Strimel, 2017). The large number of possible solutions to open-ended design problems and the elements of creativity and innovation can lead to difficulty in assessing student work using traditional approaches such as grading scales (Bartholomew, 2017; Kimbell, 2007, 2012; Pollitt, 2004, 2012). Even though rubrics, portfolios, and other assessment approaches have been recognized as potential options to assist teachers in assessing open-ended design activities (Kimbell, 2007, 2012), there are conflicting ideas and opinions regarding the best approach for assessing open-ended problems with validity, reliability, and efficiency (Pollitt, 2012). Further complicating the matter are teacher bias and subjectivity issues, which can influence assessment practices even when using rubrics and rubric-based approaches to assessment (Pollitt, 2004). Portfolios, along with student worksheets, are often used in conjunction with rubrics and criteria when evaluating design-based assignments; however, these approaches frequently require considerable time and effort for teachers to implement (Schilling & Applegate, 2012). While technology tools are increasingly being leveraged to assess students' creativity and design-thinking skills (Denson et al., 2015), these tools can include similar problems via validity, reliability, and teacher biases (Pollitt, 2012).

Adaptive Comparative Judgment

ACJ is an assessment technique based on the principle of comparative judgment that was originally developed through the work of Thurstone (1927). Thurstone argued that human comparative judgments (judgments between two items) are more valid and reliable than subjective decisions based on some predetermined quality (Pollitt, 2004; Thurstone, 1927). Pollitt (2012) revisited Thurstone's work and piloted the use of comparative judgment (CJ) for the assessment of open-ended problems (Pollitt, 2004, 2012) In doing so, Pollitt utilized assessors without a rubric to tally a score for each student's project portfolio; rather, in a CJ approach assessors simply viewed pairs of student work and identified which item was "better" based on their own professional expertise. This judgment process was repeated with different pairings until a final rank order of student work was produced. A study by Pollitt and Murray (1993) that investigated the assessment of foreign language speaking was one of the first applications of CJ on modern assessment, and the results demonstrated a high level of reliability and validity for using CJ to assess student abilities. Pollitt's continued work with CJ eventually led to ACJ—an *adaptive* version of CJ that uses an algorithm to assist in lowering the number of judgments necessary to reach an appropriate reliability level (Pollitt, 2004, 2012). The ACJ process presents items for comparison that have similar win-loss records in order to increase the reliability and efficiency of the rank order production. As with CJ, in ACJ a judgment is made holistically based on an assessor's professional expertise, experience with the subject area, and an identified "holistic statement," which frames the judgment (Pollitt, 2012). Once a "winner" is chosen out of each pairing, the system records the "win-loss" record for each item and facilitates the next comparison from the pool of items.

When employing ACJ techniques through *CompareAssess*, paired comparisons are completed until every item has been compared at least one time with another item – this is referred to as a "round" of judgment. The system updates the rank order of student work after each round so that the "winning" items rise and the "losing" items fall. The assessors (often referred to as "judges" in ACJ literature) continue to make judgments through multiple rounds to reach a desired level of reliability, which is calculated by the system after each round. Generally, the reliability of the final rank order increases as more rounds of judgments are completed (Pollitt, 2012); however, our experience suggests a point of diminishing returns where more judgments will only increase the reliability ever so slightly (experience has demonstrated that this happens after approximately 12 rounds of judgment). The final result of the ACJ process includes: (1) rank-order and parameter value statistics for all the items being compared, (2) Rasch-model misfit statistics for items and judges that can be used to identify any potentially significant areas of disagreement, and, if it is collected, (3) comments or justifications surrounding each decision made by the judges. The final rank order has been used in a variety of ways including, but not limited to the following: assigning grades, informing teacher pedagogy and student practice, as a formative tool for improvement, and as a fractional portion of total points received for a given assignment (Bartholomew, Strimel, & Yoshikawa, 2019; Bartholomew & Yoshikawa, 2018; Jones & Wheadon, 2015; McMahon & Jones, 2015).

The web-based ACJ portal used in this research, *CompareAssess* (DigitalAssess, 2017), is marketed out of England and has repeatedly demonstrated the ability to reach a high reliability level ($r > .9$) within approximately12 rounds of judgments (Bartholomew & Yoshikawa, 2018). ACJ, and *CompareAssess* specifically, have been studied and demonstrated reliable, valid, and effective results in both formative and summative assessment approaches at middle schools, high schools, and higher education levels (Bartholomew et al., 2017; Bartholomew & Yoshikawa, 2018; Hartell & Skogh, 2015; Kimbell, 2012; Pollitt, 2004; Seery et al., 2012; Strimel, Bartholomew, Jackson, Grubbs, & Bates, 2017). However, ACJ as a tool for elementary teachers to assess open-ended design activities has not been tested (Bartholomew & Yoshikawa, 2018). Additionally, ACJ as an approach and *CompareAssess* as a tool, for an *individual* teacher to use in assessment, have not been tested (Bartholomew & Yoshikawa, 2018) because an individual teacher approach runs counter to the original intent of the approach that relies on, and uses, multiple judges, Rasch modeling, and the adapted algorithm from Thurstone (1927) to calculate the reliability of the emerging rank order of items. Despite this seemingly contradictory approach to ACJ assessment we believe it is important, and useful, to understand the implications of an individual teacher employing this tool for assessment in their classroom as an individual teacher approach most closely mirrors many of the current teacher practices in assessment. Therefore, this study purposefully sought to investigate the potential of ACJ for elementary school STEM design assessment use and the possibility of an individual teacher leveraging ACJ for their own assessment of student work.

METHOD

This study took place in a small suburban school district located in the Midwestern United States. This district is composed of a mostly Caucasian (85.6%) middle-class population and serves approximately 10,000 students with a small free/reduced lunch student population (22%).

Teachers.

Following the receipt of IRB approval, four teachers (two fourth grade teachers and two Kindergarten teachers) from one elementary school in this district were recruited for participation in this study based on their interest in ACJ and their STEM integration efforts in their classrooms. Each of the teachers was recommended by the school instructional excellence coach and had similar years of experience, licensure qualifications, and interest in STEM integration. All of the teachers were Caucasian, had taught for more than five years, and had little previous experience with STEM integration. Three of the teachers were female and none of the teachers had prior experience with ACJ assessment techniques. All teachers were trained prior to the study on the STEM activities and *CompareAssess*. Throughout the study, a member of the research team was present during each class to ensure fidelity of implementation (in both the classroom activities and ACJ use). The teachers led their students through three open-ended STEM design activities which involved students working in groups to employ an elementary school level design process and resolve a posed problem from a classroom text (see Table 1).

Activities.

A total of 100 students, from the four participating classrooms, were recruited for participation in the study, which took place during three in-class time blocks (between 60 to 90 minutes each) spanning three weeks. These activities represented the first time students were involved with STEM activities during this school year. The students, in each grade level, were presented with a problem from a book currently being read in the class (e.g., *Pink and Say* by Patricia Polacco for the fourth grade students), and then they were asked how they might solve the problem (see Table 1). Students worked in groups of 2-3 (uniquely-formed for each design problem by the classroom teachers) to identify the criteria and constraints around the problem (from the book), explored pertinent questions, brainstormed ideas, and examined possible solutions. The students filled out one design worksheet (developed collaboratively by the teachers involved in this project) per group while working on the problem (see Figure 1).

Students were provided with building supplies to use for creating a mock-up of their solution to the presented problem

Table 1. Student Problems and Supporting Text by Grade

Grade	Problem	Mentor Text
K	1 – Design and build a box that does not allow frogs but allows bugs access in/out	*My Bug Boy* by Pat Blanchard
K	2 – Design and build something to help the dragonfly catch prey	*Dragonflies* by Margaret Hall
K	3 – Design and build something to help the toad come out into the sunlight without getting too hot	*Toads* by Alyse Sweeney
4	1 – Design and build something to help Pink carry Say	*Pink and Say* by Patricia Polacco
4	2 – Design and build something to conceal and carry a secret message for the army	*Great women of the Civil War* by Molly Kolpin
4	3 – Design and build something to help carry soldiers	*The Terrible, Awful Civil War* by Kay Mechiserdech

Lesson 1: 4th Grade STEM Thinking Sheet	
Criteria: • Design must lift Say off the ground • Design should be no taller or wider than 12 inches	**Constraints:** • Limited to materials provided • 1 - Dixie Cup • 2 feet - String • 16 - popsicle sticks • 1 - 9’ x 11” sheet of paper • 2 - normal length straws • Masking Tape • Limited time for creation
Ask: How can we help Pink carry Say?	**Explore:**
Model:	**Evaluate:**
Explain:	**Other ideas:**

STEM Thinking Assessment Rubric				
	Emerging	**Approaching**	**Meets Expectations**	**Exceeds Expectations**
Collaboration	Group has no product • No evidence of collaboration	Group has multiple products • No evidence of collaboration	Group has no product • Evidence of collaboration	Group has one product • Evidence of collaboration with designated roles
Process of thinking	No evidence of thinking on design process page	Attempts evidence of thinking on design process page • May not make sense	Clear evidence of thinking on design process page about the text, brainstorming, and sketch of the model	Clear evidence of thinking of the design process page with depiction of connections from brainstorming tomode (arrows, lines)
Product Creation	Limited evidence of criteria and constraints. Prototype does not fix the problem	Some evidence of criteria and constraints and prototype solves a problem • Solved a problem, but not directly connected to the text	Evidence of adherence to criteria and constraints and model solves the problem from the text	Evidence of adherence to criteria and constraints and model solves the problem from the text. Students include a written description of the produce

Figure 1: Student Worksheet and Rubric for the STEM Design Activities (4th grade, Problem 1)

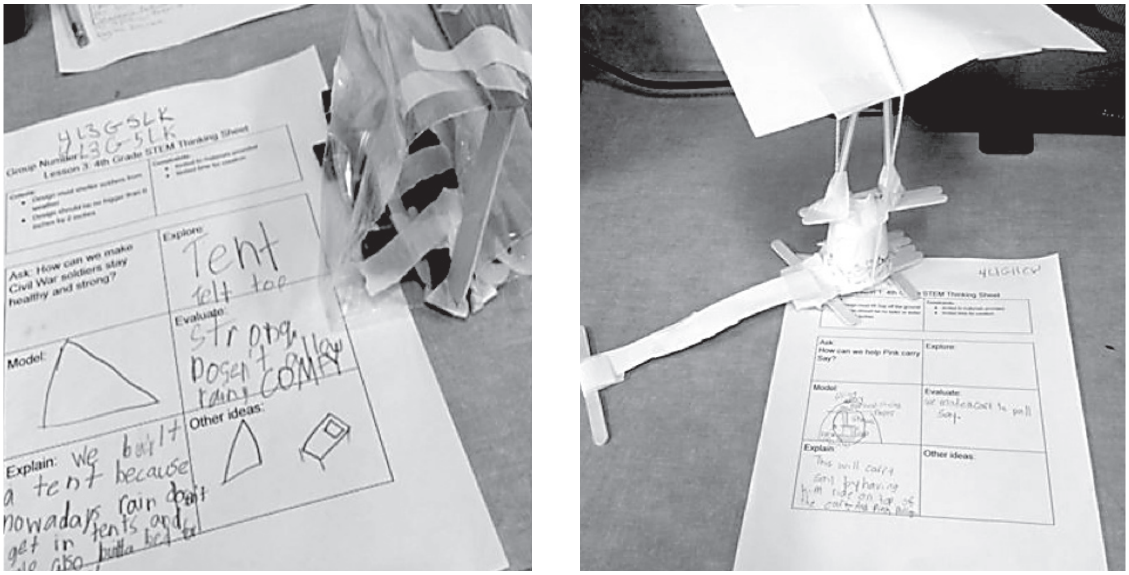

Group Number: Y13G5LK
Lesson 3: 4th Grade STEM Thinking Sheet
Ask: How can we make Civil War soldiers stay healthy and strong?
Explore: Tent
Evaluate:
Model:
Other ideas:
Explain:
Ask: How can we help Pink carry Sey?
Explore:
Evaluate:
Model:
Explain:
Other ideas:

Figure 2. Student Worksheet and Mock-up Examples (Left: 4th grade, problem 3; Right: 4th grade problem 1)

(see Figure 2). While working on the problem, students were allowed to move freely about the classroom, solicit help from neighbors or the teacher, and obtain additional supplies, if needed, with permission from the teacher. An example of the student group mock-ups and finished worksheets are shown in Figure 2.

Following each activity the student groups submitted their worksheets and mock-ups for assessment. Each student mock-up was collected and a picture of the mock-up, and the accompanying worksheet, was taken for assessment. In alignment with the research protocol, no student names were collected—rather a unique group identifier was placed on the student worksheets and used to link student work, scores, and teacher assessment practices.

Assessment.

After pictures of the student work were collected and uploaded to *CompareAssess*, the teachers were instructed to assess their students' projects twice—using both rubric-based and ACJ assessment approaches for each submission. The assessment was completed for each portfolio and prototype that represented the work for one student group. Restrictions around the viewing of student work maintained that each piece of work was only assessed by the students' teacher. However, efforts to minimize the possible influence of two specific lurking variables (e.g., maturation and history) resulted in the researchers designating a specific assessment sequence for teachers around ACJ and traditional rubric methods for assessment (see Figure 3).

	Problem 1		Problem 2		Problem 3	
Kindergarten Teacher 1	Traditional Assessment	Adaptive Comparative Judgment	Adaptive Comparative Judgment	Traditional Assessment	Traditional Assessment	Adaptive Comparative Judgment
Kindergarten Teacher 1	Adaptive Comparative Judgment	Traditional Assessment	Traditional Assessment	Adaptive Comparative Judgment	Adaptive Comparative Judgment	Traditional Assessment
4th Grade Teacher 1	Traditional Assessment	Adaptive Comparative Judgment	Adaptive Comparative Judgment	Traditional Assessment	Traditional Assessment	Adaptive Comparative Judgment
4th Grade Teacher 1	Adaptive Comparative Judgment	Traditional Assessment	Traditional Assessment	Adaptive Comparative Judgment	Adaptive Comparative Judgment	Traditional Assessment

FIGURE 3. Assessment Approach Sequence for Teachers

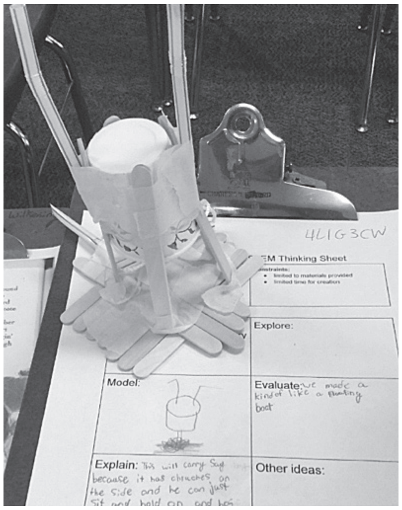
Teachers followed the specified approach (Figure 3), completing both the traditional assessment, using the rubric in Figure 2, and an additional web-based ACJ assessment through the *CompareAssess* portal. The ACJ portal facilitated the assessment by prompting teachers to login and then select a project for assessment (e.g., Project 1-3) after which teachers were then shown two images, each containing both the student's worksheet and mock-up, and asked to make a comparative judgment as to which was better. Teachers were instructed to make the decision between the two items holistically during ACJ assessment while also bearing specifically in mind the rubric and assignment description (see Kimbell [2007] and Pollitt [2012] for a thorough explanation and rationale behind the holistic nature of ACJ assessment). Teacher assessment times were collected automatically through *CompareAssess* and electronically through a web-based timer for the rubric-based approach.

Data Collection.

Data was collected to investigate the overarching research question around the utility of ACJ, as an assessment tool, for STEM design activities – as performed by an individual teacher (see Figure 4). In order to compare the workload of the teachers for each of the assessment approaches, the time teachers spent assessing the student work, in each of the two approaches, was collected. Additionally, the results from the ACJ sessions were gathered and the teacher perceptions of ACJ, both as an assessment tool overall and specifically as a possible option for individual classroom teacher use, were collected through a post-study questionnaire. These responses were specifically gathered and analyzed to better understand the teacher experiences with, and perceptions of, ACJ for individual assessment and classroom use.

Traditional Scoring Rubric

	1. Emerging	2. Approaching	3. Meets expectations	4. Exceeds expectations
Collaboration	Group has no product and no evidence of collaboration	Group has multiple products but no evidence of collaboration	Group has one product and there is evidence of collaboration	Group has one product and evidence of collaboration with designated roles
Process of thinking	No evidence of of thinking on design process page	Attempts evidence of thinking on design process page but may not make sense	Clear evidence of thinking on design process page through text, brainstorming, and sketch of model	Clear evidence of thinking on design process page with depiction of connections from brainstorming to model (e.g., arrows, lines)
Product Creation	Limited evidence of criteria and constraints. Prototypes does not fix the problem	Some evidence of criteria and constraints and prototype solves a problem but not directly connected to text	Evidence of criteria and contraints and model solves problem from text	Evidence of criteria and constraints and model solves the problem from the text. Students include a written description of the product



Group Rank and Score

Group	ACJ Rank	Tradiational Assesment score
4	1	10
1	2	11
10	3	9
7	4	9
9	5	8
5	6	9
8	7	8
2	8	8
3	9	9

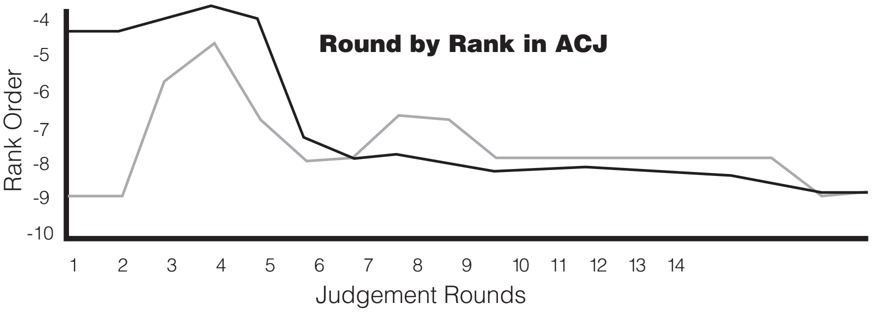


FIGURE 4. Assessment Data for Teacher 4, Group 3, Project 1

These data were collected in an effort to explore the teacher workload, their perceptions of using ACJ for assessment, and the similarities and differences between the two different approaches to grading. All data were collected, conditioned, and analyzed using *SPSS* statistical software (Version 23). The data collection and the accompanying analysis for each research question are presented here.

FINDINGS

The findings from this study will be presented here in conjunction with each of the specific research questions that framed this work.

RQ_1: What relationship exists, if any, between student achie vement measures obtained through ACJ and rubric-based assessment approaches for elementary school STEM design activities?

The first research question investigated the potential relationship between ACJ and traditional assessment approaches for the elementary school STEM design activities. The teacher assessment scores and ranks—obtained through rubric-based approaches and *CompareAssess*—were collected, and a Spearman correlation test was conducted for each problem by the teacher (see Table 2).

First, it should be noted that a negative correlation was expected as a lower rank corresponds with a higher-quality item – thus a negative correlation demonstrates alignment between the two approaches while a positive correlation would suggest dissonance. Second, it should be noted that the traditional rubric was created with the cooperating teachers and the teacher marks are based on their knowledge, understanding, and expectations for the associated grade level. Therefore, while it stands to reason that fourth grade students could be expected to outperform Kindergarten students on a given activity based on their maturity level and experience, each rubric used by teachers was designed with student's age, abilities, and backgrounds in mind.

Interestingly, two of the teachers (1, 4) were significantly aligned in their ACJ judgments and their traditional assessment for the first two problems, but they were not significantly aligned in problem 3. Teacher 3's assessment practices were significantly correlated with the ACJ rank for the second assignment but were not significantly correlated for the other two problems. Finally, teacher 2 never demonstrated a significant correlation between the ACJ-ranks received and the score received by students through traditional assessment. A closer investigation of Teacher 2 revealed that almost every group in Teacher 2's classroom received the same grade for each problem through traditional assessment; for example, on problem 2 every group received a score of "6" except for one group, which was awarded a "5." Thus, these correlations were likely not significant as there was very little variation in score received through traditional assessment while the ACJ ranking demonstrated a spread from 1-7.

Table 2. Correlation between ACJ rank and Rubric Grade Received for Each Assignment

Teacher	Problem r, sig. (2-tailed)		
	1	2	3
Teacher 1 (Kindergarten, n = 7)	-.94*, .00	-.79*, .02	-.52, .19
Teacher 2 (Kindergarten, n = 7)	-.68, .10	-.58, .13	-.67, .07
Teacher 3 (Fourth Grade, n = 7)	-.12, .80	-.88*, .02	-.52, .30
Teacher 4 (Fourth Grade, n = 9)	-.67*, .05	-.80*, .01	-.47, .20

*Correlation is significant at the 0.05 level (2-tailed)

RQ$_2$: What are the implications for teacher workload associated with ACJ and rubric-based assessment approaches to an elementary school STEM design activity?

The second research question emphasized the similarities and differences in the workload required for teachers to use each assessment approach (rubric-based vs. ACJ)—as measured through the overall time taken for each form of assessment. The total time used for traditional assessment and ACJ was recorded and a one-sample *t*-test was conducted to compare the differences in the total time taken by teachers using traditional assessment approaches and ACJ assessment approaches. There was a significant difference in the total time taken for teachers to conduct the assessment of each project through traditional (*M* = 389 seconds, *SD* = 152.14) and ACJ assessment approaches (*M* = 715 seconds, *SD* = 323.52) approaches, *t*(11) = -2.98, *p* = .01.

Closer analysis of the data revealed two significant outliers (more than triple the time taken in any other assessment) in the time taken by Teacher 4 to complete the ACJ assessment for projects 1 and 3. These outliers, which may have resulted from a variety of factors (e.g., the teacher stepped away from the computer while the ACJ session was open and the extended time was recorded), were removed and the one-sample *t*-test was conducted again to further investigate this relationship. The one-sample *t*-test, conducted with the outliers removed, once again demonstrated a significant difference in the time taken for teachers to conduct the assessment through traditional (*M* = 412 seconds, *SD* = 153.15) and ACJ (*M* = 588 seconds, *SD* = 131.38) approaches, *t*(9) = -3.95, *p* = .006. These results indicate that ACJ, as an assessment approach, took significantly longer for the teachers in this study than the traditional assessment approaches.

In order to investigate the potentially significant influence of teacher, grade-level, or assignment on the total time for each assessment approach, we used a three-way main effect model with nesting. Following a check to ensure the required assumptions were met, the analysis was conducted and revealed that there was no significant influence on the time taken in either assessment approach from teacher, grade-level, or assignment.

To further investigate the quantitative findings from this research, the teachers were asked, as part of the post-study questionnaire, how the use of ACJ for assessment compared with traditional approaches to assessment in terms of teacher workload. Overall, the teachers were undecided about whether ACJ took more, less, or the same amount of time as traditional methods of assessment; of the four teachers surveyed, half (2) responded that ACJ took less time than traditional forms of assessment, one teacher marked that it took same amount of time, and one teacher marked that ACJ took more time. When the teachers were asked for further clarification they responded with comments that suggested that ACJ was more time-intensive than traditional approaches. A few qualitative comments from the teachers include:

> *"[ACJ] took more time than I originally thought it would"*
>
> *"I wish ACJ was a bit more efficient. I felt like there were more steps than were necessary while assessing projects"*
>
> *"At first [ACJ] seemed to take a long time. It seemed like it got faster as I did the evaluations more"*
>
> *"Getting [the ACJ done] took a considerable amount of time. I think the traditional rubric was slightly easier and teachers would need to understand the benefits of comparing projects in order to see its value"*

RQ$_3$: What are the practical implications of implementing ACJ assessment for elementary school STEM design activities by individual teachers?

The third research question investigated the potential possibility of an individual classroom teacher using ACJ for assessment of student work in order to address the lack of research done in this area (Bartholomew & Yoshikawa, 2018; M. Wingfield, personal communication, May 17, 2017). This was important as the majority of classroom assessment, in the current educational culture in the United States, involves a teacher assessing student work individually and then assigning grades. The data collected around this question comes from two sources: post-study questionnaires completed by the teachers and the round-by-round exploratory analysis of rank-order for the ACJ sessions completed by the teachers.

The post-study questionnaire sought to specifically elicit teacher perceptions of ACJ, as an approach, and *CompareAssess*, as a tool, for an individual teacher to use in assessment, as compared with traditional rubric-based approaches to assessment. Recognizing the small sample size ($N = 4$), we position the teacher responses as informative and stimulating in terms of guiding future research around ACJ. The findings from the post-study questionnaire, classified by theme, are presented next:

Ease of Use

Overall the teachers believed that ACJ was easier to use for assessment than rubric-based forms of assessment with 75% (3 out of 4) teachers noting that ACJ was easier to use than a rubric for assessment, whereas one teacher marked that it was the same, in terms of difficulty. When asked about the ease of use several teachers commented:

> *It's easy to compare two images side by side, I can zoom in, I see the same project more than once, compared to other items, I can add comments.*
>
> *I liked that ACJ made it easy to see projects side by side in a comparison. It made it easy to see the differences in the projects, and assess the projects accordingly. I liked the speed with which I could assess projects with ACJ...it was faster than going through the rubric*.

Confidence in Results.

All of the surveyed teachers responded that they were "confident in the results obtained from ACJ" when surveyed. Every teacher identified the same confidence level in the results obtained from ACJ as that from traditional assessment approaches. Related, all teachers marked that the rank orders from ACJ were "similar in usefulness" to traditional assessment results. Teachers suggested using the results as a learning tool for students (i.e., using the top ranked item as an example for discussion) or as a portion of the student's final grade from an assignment. When asked about their confidence one teacher remarked:

> *The same project continued to pop up as the best one so I was confident in my decision.*

Similarly, when asked about the usefulness of ACJ and possible future uses two teachers noted the potential for future use of ACJ in their classrooms:

> *[ACJ] is quick to use and has lots of use for what we do in 4th grade.*
>
> *I would like to try this with writing.*

Implications of Individual Teacher Use

While previous work with ACJ has revolved around groups of judges completing the judgments, this study emphasized the exploration of the potential for using ACJ with one teacher. In order to do this, we sought to investigate the implications of this use and explore the number of judgment rounds necessary for an *individual* teacher to use ACJ effectively in the classroom and obtain a useful rank order. In previous work with ACJ, and *CompareAssess*, the resulting reliability level of the rank order has been a tool for identifying a "solid" rank (Pollitt, 2012), however, with only one individual performing the ACJ assessments in each session the reliability level was no longer a useful measurement around which to determine the stopping point for judgments (M. Wingfield, personal communication, May 17, 2017).

Fully recognizing limitations in our approach; namely, not relying on/trusting the computed reliability with the ACJ-based statistics, and not using the approach (ACJ) or the interface (*CompareAssess*) as intended (i.e., with a group of judges)—we informed the teachers to continue with ACJ judgments until they had completed at least eight rounds of judgment for each of the assignments. This was done intentionally to facilitate both random and adaptive pairings—the first five rounds of judgment displayed random pairings while the later rounds included the "adaptive" component with similarly judged items displayed in pairs. The rank orders of student work for each round were compared; round by round, to identify how the compared items moved between rounds based on the judgments made by teachers. Additionally, the eventual top-ranking item (from round 8) was tracked in an effort to explore how this item's rank fluctuated through the rounds of judgment. Figure 5 presents the findings from this exploratory exercise with the rank order at the conclusion of round 8 identified in gray and the top-ranking item identified with black.

Teacher 1, Project 1

		ROUND							
		1	2	3	4	5	6	7	8
RANK	1	B	B	B	B	B	B	A	A
	2	D	A	H	D	A	A	B	B
	3	C	H	F	A	H	H	H	H
	4	A	F	A	C	F	D	D	D
	5	H	D	D	F	D	C	C	F
	6	F	C	C	H	C	F	F	C
	7	G	G	G	G	G	G	G	G

Teacher 1, Project 2

		ROUND										
		1	2	3	4	5	6	7	8	9	10	11
RANK	1	H	H	H	H	H	H	H	H	H	H	H
	2	D	E	B	F	F	F	F	F	F	F	F
	3	E	F	F	E	D	B	B	C	C	C	C
	4	F	D	E	D	C	C	C	B	B	B	B
	5	G	C	G	B	B	D	D	D	D	D	D
	6	B	B	C	C	E	E	E	E	E	E	E
	7	C	G	D	G	G	G	G	G	G	G	G
	8	A	A	A	A	A	A	A	A	A	A	A

Teacher 1, Project 3

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	H	H	H	H	H	H	H	H	H	H
	2	F	A	D	D	D	D	A	A	A	A
	3	C	F	E	E	G	G	D	D	D	D
	4	A	E	A	C	A	G	G	G	E	E
	5	E	C	G	A	E	E	E	E	G	G
	6	B	D	F	G	C	C	C	C	C	C
	7	D	B	C	F	F	B	B	F	F	F
	8	G	G	B	B	B	F	F	B	B	B

Teacher 2, Project 1

		ROUND							
		1	2	3	4	5	6	7	8
RANK	1	C	E	E	B	B	B	B	B
	2	B	B	B	F	C	E	E	E
	3	E	F	C	C	E	F	F	F
	4	A	C	F	E	F	C	C	C
	5	F	A	G	A	A	A	A	A
	6	D	G	A	G	D	G	G	G
	7	G	D	D	D	G	D	D	D

Teacher 2, Project 2

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	C	F	E	E	E	E	E	E	E	E
	2	F	E	G	D	D	D	D	D	D	D
	3	E	G	D	F	A	A	A	A	A	A
	4	A	D	F	H	H	F	F	F	H	H
	5	G	A	H	G	G	H	H	H	F	F
	6	B	C	C	A	F	B	B	B	B	B
	7	D	B	A	C	C	G	G	G	G	G
	8	H	H	B	B	B	C	C	C	C	C

Teacher 2, Project 3

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	B	E	E	E	E	E	E	E	E	E
	2	J	J	J	A	B	A	A	B	B	B
	3	E	F	A	B	J	B	B	A	A	A
	4	F	G	F	D	A	J	J	J	D	F
	5	D	D	B	F	K	F	F	F	F	D
	6	A	A	K	J	F	D	D	D	J	K
	7	C	B	D	H	D	K	K	K	K	J
	8	G	H	G	C	C	C	C	C	C	C
	9	K	K	C	G	G	H	H	H	H	H
	10	H	C	H	K	H	G	G	G	G	G

Teacher 3, Project 1

		ROUND							
		1	2	3	4	5	6	7	8
RANK	1	D	C	G	C	C	C	C	C
	2	G	D	C	G	D	D	D	D
	3	C	A	D	D	E	G	G	G
	4	A	F	A	A	G	E	E	E
	5	B	G	F	E	A	A	A	A
	6	E	E	E	F	F	F	F	F
	7	F	B	B	B	B	B	B	B

Teacher 3, Project 2

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	E	A	E	E	D	D	D	D	D	D
	2	F	E	D	D	E	E	E	E	E	E
	3	A	F	A	F	A	A	A	A	A	A
	4	D	D	C	A	C	F	F	F	F	F
	5	C	C	F	C	F	C	C	C	C	C
	6	G	G	G	G	G	G	G	G	G	G

Teacher 3, Project 3

		ROUND							
		1	2	3	4	5	6	7	8
RANK	1	F	G	D	D	D	D	D	D
	2	G	D	F	G	E	E	E	E
	3	D	F	E	E	B	A	A	A
	4	A	E	G	F	A	G	G	G
	5	C	A	B	A	G	B	B	B
	6	B	B	A	B	F	F	F	F
	7	E	C	C	C	C	C	C	C

Teacher 4, Project 1

		ROUND											
		1	2	3	4	5	6	7	8	9	10	11	12
RANK	1	F	B	B	H	H	H	H	H	B	B	B	B
	2	H	H	H	F	B	B	B	B	H	H	H	E
	3	G	F	F	E	F	E	E	E	E	E	E	H
	4	B	G	F	B	F	F	F	F	F	F	F	F
	5	E	A	E	C	G	G	G	G	F	F	F	F
	6	A	F	C	F	E	F	F	F	G	G	G	G
	7	F	A	E	G	C	D	D	C	C	D	D	D
	8	D	D	G	A	A	C	C	D	D	C	C	C
	9	C	C	D	D	D	A	A	A	A	A	A	A

Teacher 4, Project 2

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	D	H	E	E	F	F	F	F	E	E
	2	F	J	H	J	J	E	H	H	F	F
	3	J	E	J	H	E	H	E	E	H	H
	4	B	F	F	F	H	J	J	J	J	J
	5	E	D	D	D	D	D	D	D	D	D
	6	G	B	G	A	C	A	A	A	A	A
	7	A	A	B	B	A	C	C	B	B	C
	8	C	G	A	G	B	B	B	C	C	B
	9	H	C	C	C	C	G	G	G	G	G

Teacher 4, Project 3

		ROUND									
		1	2	3	4	5	6	7	8	9	10
RANK	1	A	E	J	J	J	J	J	J	J	J
	2	D	J	E	E	D	E	E	E	E	E
	3	J	A	C	D	C	D	D	D	D	D
	4	E	F	D	A	E	A	A	A	A	A
	5	F	D	A	C	A	C	C	C	C	C
	6	C	G	F	B	B	F	B	B	H	H
	7	G	C	G	F	F	H	H	H	B	B
	8	B	H	B	G	H	B	F	F	F	F
	9	H	B	H	H	G	G	G	G	G	G

FIGURE 5. Round by Round Rank of Student Work for Each Teacher and Project Assignment

The analysis showed that, when completed individually, the number of rounds required to reach a "stable" rank order may be, at least partially, contingent on the judging style of each teacher/judge. Holistically, our analysis suggested that somewhere between six and ten rounds of judgment the rank order of items began to stabilize when ACJ was completed by an individual judge. For some teachers (Teacher 3) six to eight rounds of judgment appeared to consistently produce a more "stable" rank order, whereas for other teachers (Teacher 4) ten rounds of judgment was still not enough for the rankings of student work to demonstrate consistency. We fully recognize that a variety of factors influenced these rank orders and the preliminary indications including, but not limited to, the number of pieces of student work, the type of work being assessed, the teachers involved in this study and their training, experiences, grading practices, background, and exposure to ACJ. Further, it should be noted that these findings—which are exploratory in nature—were contingent on the ACJ-platform (*CompareAssess*), the order items were presented to judges, and the way items were paired. Additionally, while these findings are confined to the judges in this study, the student work, and the ACJ tool used for this research, the analyses, findings, and implications are important and provocative in terms of future implications and research around ACJ – especially if classroom teachers continue to complete assessment for student work individually.

DISCUSSION AND CONCLUSION

Similar to previous research with high school (Pollitt & Crisp, 2004; Newhouse, 2011; Steedle & Ferrara, 2016), middle school (Bartholomew, Reeve, Veon, Goodridge, Stewardson, Lee, & Nadelson, 2017; Bartholomew, et al., 2017), and post-secondary students (Seery, Canty, & Phelan, 2012; Strimel et al., 2017), the correlation between the ACJ ranking and the students' scores obtained through traditional assessment approaches at the elementary school level was significant for select teachers during certain problems. However, for other problems these same teachers, and the other teachers involved, did not demonstrate significant correlations between their traditional assessment practices and the rank order of student work. This suggests that alignment of ACJ results with traditional forms of assessment may not be universal and may be a function of a variety of factors such as grade level, problem scenario, number of judges, and teacher assessment strategies and practices. Additionally, it should be considered that although ACJ provides a rank order of the included items, it does not speak to the overall, or specific, quality of the items (e.g., the top-ranked item may still not be very "good" in terms of functionality of teachers' expectations, or the lowest ranked items may be considered acceptable work according to the assignment criteria). It is possible that neither approach to assessment was truly a valid measure of student achievement or learning. It is also possible that only one of the assessment approaches is valid, whereas the other approach is not. We also wish to draw attention to the fact that no reliability or validity data was available for the teacher-created rubrics—these are important considerations, which may influence areas for future research and exploration.

Our findings, while limited in nature by the small sample of teachers, the problem context, and the research design, highlighted significant differences in assessment approaches between teachers. Although the provided rubric guided teachers to use their expectations of students in assessment, it was apparent from the results that these expectations were sometimes very different for different teachers. Also, while some teachers traditional assessment scores varied significantly, other teachers had little to no variation in the scores students were given through traditional assessment; a difference that was especially highlighted in comparison with ACJ because ACJ systematically established differences between each student through the ranking process. It was evident, from our findings and observations, that some of the teachers included in this study routinely had very little deviation in the scores assigned to students with many, if not all, students receiving full marks for simply *completing* an assignment. We also wish to point out that, keeping in line with commonly practiced approaches, no reliability measures were attempted in relation to the traditional scoring approaches utilized by the teachers. The teachers included in this study used rubrics with reliability testing only in "high-stakes" test scenarios (i.e., state- and nationally-administered tests).

A look at the teacher comments and the collected time records revealed that ACJ took significantly longer than traditional forms of assessment. The findings from this study, uniquely situated with an individual teacher using ACJ, align with previous research (Pollitt, 2012) around groups of judges using ACJ. The significant difference in time was consistent for all teachers in this study and teacher comments on the post-study questionnaire supported this sentiment. Despite the increased time required to implement ACJ, teachers in this study were positive toward its potential for integration into their classrooms and believed ACJ was easier in terms of making judgments than making criteria score decisions using traditional forms of assessment. Teachers expressed confidence in the results obtained through ACJ and the majority of the suggestions for improvement were focused on the interface of the software platform rather than the actual approach.

While ACJ may take more time than traditional forms of assessment, several benefits of ACJ, which have also been identified elsewhere, were identified by teachers in this study including: the comparability of the results obtained through ACJ and traditional assessment practices (Strimel et al., 2017), the ease of implementation in classrooms (Bartholomew, 2017), a holistic emphasis in assessment (Kimbell, 2007), and the prospects of using ACJ for student learning (Bartholomew et al., 2017). While the teachers in this study recognized that ACJ took more time than traditional assessment, several pointed out that the built-in feedback and comparison function of *CompareAssess* was an added benefit that may actually work to expedite the process of assessment in certain settings. Therefore, while our findings indicate that traditional assessment approaches were more time efficient, there may be several other important factors (i.e., the time required to provide feedback to students), which were not taken into consideration.

Based on our findings we contend that ACJ is most feasible used as directed – with multiple judges. Although this runs counter to commonly practiced educational assessment tactics in K-12 classrooms, this method appears to not only be the most valid and reliable but also to be the most effective and efficient. We recommend that future research into the possibilities and implications of individual teacher use be conducted to investigate possible widespread implementation of ACJ by individual teachers and other potential models which may increase the validity, reliability, utility, and efficiency of its integration. For example, comparing student work with items on a known scale/rank could potentially be useful in terms of facilitating judgments by an individual teacher while still collecting reliability and validity measures.

Our findings revealed that the number of judgment rounds required to reach a “stable” rank order was different for each teacher/judge. This is sensible given the variety of difference in teacher perceptions, backgrounds, and the factors involved in assessment of student work (Alkharusi, 2011; Crossman, 2004; Dietrich, 2010; Rice, 2010). From our research, we identified a range of 6-10 rounds of judgment as a potential basis for future research, practice, and implementation of ACJ by an individual.

Despite the differences in the number of rounds required for a stable rank to appear, the teachers hinted at a potential increased efficiency at identifying the top-ranking item over the course of the research (spanning three design projects) suggesting that, with time, teachers may become more efficient at producing a stable rank order and identifying the relative quality of student work through ACJ. Future research could focus on identifying the number of *required* rounds for a stable rank to appear and the potential for teachers to increase in judgment efficiency over time. Additionally, the possibility of using ACJ to assist individuals and teams of teachers in reducing inherent teacher biases (Bartholomew, 2017) and implementing different approaches to assessment, merits further investigation, research, and discussion.

***Scott R. Bartholomew, Ph.D.**, is an assistant professor of Engineering/Technology Teacher Education at Purdue University, West Lafayette, Indiana.*

***Greg J. Strimel, Ph.D.**, is an assistant professor of technology leadership and innovation at Purdue University, West Lafayette, Indiana.*

***Liwei Zhang** is a masters student and graduate research assistant in the Engineering Technology Teacher Education program at Purdue University.*

***Jessica Homan** is a library media specialist at Noble Crossing Elementary. Jessica is a Project Lead the Way lead teacher, technology lead teacher, and inquiry lead teacher for Noblesville Schools, Indiana.*

REFERENCES

Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal of Turkish Science Education,* 8(2), 39-48.

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science aspirations and family habitus: How families shape children's engagement and identification with science. *American Educational Research Journal*, 49(5), 881–908.

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

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H., & Reeve, E. M. (2018). 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., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., Nadelson, L. (2017). Mobile devices, self-directed learning, and achievement in Technology and Engineering Education classrooms during a STEM activity. *Journal of Technology Education*, 29(1), 2-24.

Bartholomew, S. R., & Strimel, G. J. (2017, March). The problem with assessing open-ended problems, *Techniques*. 44-49.

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school open-ended design challenge. *International Journal of Technology & Design Education*, 29(2), 363-385, https://doi.org/10.1007/s10798-018-9442-7

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

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher*, 70(1), 30-35.

Crossman, J. (2004). Factors influencing the assessment perceptions of training teachers. *International Education Journal*, 5(4), 582-590

Daugherty, M. K., Carter, V., & Swagerty, L. (2014). Elementary STEM education: The future for technology and engineering education? *Journal of STEM Teacher Education*, 49(1), 44-55.

DeJarnette, N. K. (2012). America's children: Providing early exposure to STEM (science, technology, engineering and math) initiatives. *Education*, 133(1), 77–84.

Denson, C. D., Buelin, J. K., Lammi, M. D., & D'Amico, S. (2015). Developing instrumentation for assessing creativity in engineering design. *Journal of Technology Education*, 27(1), 23-40.

Diefes-Dux, H. A., Moore, T., Zawojewski, J., Imbrie, P. K., & Follman, D. (2004). A framework for posing open-ended engineering problems: Model-eliciting activities. Paper presented at the IEEE Frontiers in Education 2004 Annual Conference.

Dietrich, C. (2010). Decision making: Factors that influence decision making, heuristics used, and decision outcomes. *Student Pulse*, 2(2), pp. 1-3.

DigitalAssess. (2017). *What we do*. Retrieved on October 20, 2017 from: http://digitalassess.com/what-we-do/#compareassess

Epstein, D., & Miller, R. (2011). Elementary school teachers and the crisis in STEM education. *The Education Digest*, 77(1), 4-10.

Hartell, E., & Skogh, I. B. (2015). Criteria for success: A study of primary technology teachers' assessment of digital portfolios. *Australasian Journal of Technology Education*, 2(1), 2-17.

International Technology and Engineering Educators Association (ITEA/ITEEA). (2000/2002/2007). *Standards for technological literacy: Content for the study of technology*. Reston, VA: Author.

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

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

Kimbell, R. (2012). The origins and underpinning principles of e-scape. *International Journal of Technology & Design Education*, 22, 123-134.

Kuenzi, J. J. (2008). Science, technology, engineering, and mathematics (STEM) education: Background, federal policy, and legislative action. *Congressional Research Service Reports*. Retrieved from http://digitalcommons.unl.edu/ crsdocs/35.

Laboy-Rush, D. (2011). Integrated STEM education through project-based learning. Retrieved from https://pdfs.semanticscholar.org/a51b/9bab3eb593b36098bf93da0d34caae927228.pdf

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3), 368-389.

Murphy, T. (2011, August 29). STEM education—It's elementary. *US News and World Report*. Retrieved from http://www.usnews.com/news/articles/ 2011/08/29/stem-education--its-elementary.

Nadelson, L. S., Callahan, J., Pyke, P., Hay, A., Dance, M., & Pfiester, J. (2013). Teacher STEM perception and preparation: Inquiry-based STEM professional development for elementary teachers. *The Journal of Educational Research*, 106(2), 157-168.

National Academy of Engineering (NAE) & National Research Council (NRC). (2014). *STEM integration in K–12 education: Status, prospects, and an agenda for research*. Washington, DC: National Academies Press.

Newhouse, P. (2011). Comparative pairs marking supports authentic assessment of practical performance within constructivist learning environments. In *Applications of Rasch measurement in learning environments research* (pp. 141-180). SensePublishers.

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

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

Pollitt, A. (2004). Let's Stop Marking Exams. Retrieved from http://www.cambridge assessment.org.uk/images/109719-let-s-stop-marking-exams.pdf

Pollitt, A., & Crisp, V. (2004). Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions? Retrieved from www.leeds.ac.uk/educol/documents/00003731.htm

Pollitt, A, & Murray, N. J. (1993). What raters really pay attention to. *Language Testing Research Colloquium*, Cambridge. Republished in Milanovic, M. & Saville, N. (Eds.), Studies in Language Testing 3: Performance Testing, Cognition and Assessment, Cambridge University Press, Cambridge.

Reeve, E. M. (2015). STEM thinking! *Technology and Engineering Teacher*, 75(4), 8-16.

Rice, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications. *Brief*, 11. Washington, DC: Center for Analysis of Longitudinal Data in Education Research.

Schilling, K., & Applegate, R. (2012). Best methods for evaluating educational impact: A comparison of the efficacy of commonly used measures of library instruction. *Journal of the Medical Library Association*, 100(4), 258-269.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, 22(2), 205-226.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29(3), 211-223.

Strimel, G. J., Bartholomew, S. R., Jackson, A., Grubbs, M., & Bates, D. G. M. (2017). Evaluating freshman engineering design projects using adaptive comparative judgment. Paper presented at the *American Society of Engineering Education 124th Annual Conference & Exposition*, Columbus, OH.

Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Educatio*n, 13(1), 16-27.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.