Evaluating Freshman Engineering Design Projects Using Adaptive Comparative Judgment

Dr. Greg J. Strimel, Purdue Polytechnic Institute

Dr. Greg J. Strimel is an assistant professor of engineering/technology teacher education in the Purdue Polytechnic Institute at Purdue University in West Lafayette, Indiana. His prior teaching experience includes serving as a high school engineering/technology teacher and a teaching assistant professor within the College of Engineering & Mineral Resources at West Virginia University.

Dr. Scott R. Bartholomew, Purdue University

My interests revolve around adaptive comparative judgment, integrated STEM learning, Technology & Engineering Design learning, and self-directed learning. I have taught at the middle-school, high school, and collegiate levels and am dedicated to strengthening Technology & Engineering Education.

Mr. Andrew Jackson, Purdue Polytechnic Institute

Andrew Jackson is currently pursuing a PhD in Technology through Purdue's Polytechnic Institute, with an emphasis on Engineering and Technology Teacher Education. His research interests are engineering self-efficacy, motivation, and decision making. Andrew is the recipient of a 2015 Ross Fellowship from Purdue University and has been recognized as a 21st Century Fellow by the International Technology and Engineering Educators Association. He completed his Master of Science in Technology Leadership and Innovation at Purdue University with a thesis investigating middle school engineering self-efficacy beliefs. He previously taught middle school and undergraduate technology courses, accompanying both experiences with classroom research to improve practice.

Dr. Michael Grubbs, Baltimore County Public Schools

Previous to my current position as Supervisor of Technology, Engineering, and Manufacturing Education of Baltimore County Public Schools, I was a Virginia Tech GRA and educator in Clayton County Public Schools.

Mr. Daniel Gordon Mendiola Bates, North Carolina State University

PhD candidate in STEM Ed - Technology, Engineering, and Design (TED) at NC State University. Research interests include Engineering mindset, model-based reasoning, computational thinking in TED, and entrepreneurial influence in TED education. 4 years k-12 teaching experience.

Evaluating Freshman Engineering Design Projects Using Adaptive Comparative Judgment

Abstract

This evidence-based practice paper examines the use of an alternative form of assessment for engineering design projects called adaptive comparative judgment (ACJ). The authors employed ACJ to assess undergraduate engineering student design projects and compared the results to traditional marking assessment techniques. The authors sought to examine reliability and validity of ACJ in comparison with traditional assessment techniques. Student work from 16 first-year engineering majors was initially graded by the course instructor and then a panel of five judges completed the ACJ method to evaluate the same work. This work consisted of design portfolios and pictures of design prototypes. The authors conducted an analysis of the reliability and validity of the ACJ when compared to the performance data of each student's prototype and the traditional rubric used by the course instructor to evaluate the project. This paper aims to introduce the method of using ACJ for engineering design projects and make the case for this method based on current research efforts and the preliminary findings of this study.

Introduction

Engineering programs are often the home of multiple open-ended student design projects. The common method of assessment for most design projects is using a predetermined rubric to assign scores to student work (Pollitt, 2004). These assigned scores can be holistic in nature or based on micro-judgments that are summated to produce a macro-judgment of student performance (Pollitt, 2004; Kimbell, 2012). However, a problematic issue with the traditional scoring of student design work using rubrics is the low reliability resulting when multiple graders assess student work (Pollitt, 2004, 2012). As a solution to this issue, Pollitt (2004) presented an alternative form assessment known as adaptive comparative judgment (ACJ). ACJ is a form of assessment that relies on comparisons of student work rather than rubrics. Bartholomew, Reeve, Veon, Goodridge, Stewardson, Lee, & Nadelson (in press) explained this process as a method of displaying a piece of work or artifact from two different students or student groups (e.g., essays, pictures, technical drawings, engineering notebooks, or design portfolios) to judges with the direction to rate which piece of work is better. The judges are not asked to provide a grade for each piece of work but rather asked to provide a holistic decision as to which artifact is better based on their own professional opinion and a provided rubric. In each round of judgment one artifact is compared to another and a decision is made by the judge as to which item is better. Rounds of judgment are conducted until a sufficient reliability level is reached and a final rank-order for student work is obtained. While some may argue against the idea of comparing students to one another, Kimbell (2012) and Pollitt (2004) explain that any kind of assessment is essentially a comparison of one thing to another. As Pollitt states, "All judgments are relative. When we try to judge a performance against grade descriptors we are imagining or remembering other performances and comparing new performances to them" (2004, p. 6).

The ACJ method of assessment has proven to be more reliable and valid than the traditional methods of assessment (Bartholomew, et al., in press; Kimbell 2012; Pollitt, 2004, 2006, 2012). The theoretical development of the ACJ assessment method has led to the formation of a grading engine by TAG Assessment titled *CompareAssess*. This product provides a platform for student work to be more easily rated by multiple judges and algorithmically

outputs the rank-order and standardized score of relative work quality. This paper will examine the use of *CompareAssess* as a means for evaluating engineering design projects of undergraduate engineering students by using multiple judges to compare the design artifacts of 16 undergraduate engineering students. The authors will analyze the reliability and validity of this method when compared to the performance data of each student's solution and the traditional rubric used to evaluate the project.

Statement of the Problem

While engineering students are often provided the opportunity to learn key concepts, hone engineering design skills, and demonstrate competencies through resolving authentic problems, it can be challenging to efficiently assess student design work to a level that is reliable and valid. However, with greater insight to alternative forms of assessments, engineering educators may be better equipped to manage the difficulties in assessing students' abilities in effectively resolving engineering design challenges. Therefore, the purpose of this research was to examine the use of the ACJ assessment approach in contrast to a traditional rubric in evaluating first-year engineering design projects.

Research Questions

The research questions that guided this study included the following:

RQ_1: What relationship, if any exists, between the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment?

RQ_2: What relationship, if any exists, between the performance of student design projects and the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment?

Engineering Design Projects, Assessment, and Adaptive Comparative Judgment

Engineering design projects have been a hallmark of engineering education for years (Dutson, Todd, Magleby, & Sorensen, 1997; Katehi, Pearson, & Feder, 2009; Cunningham, 2009). Traditionally students, presented with an open-ended problem, work in groups to solve, test, modify, and present their solution to an open-ended scenario or problem. Engineering design problems are not confined to higher education as a host of K-12 engineering initiatives, curriculum, and projects have emerged in recent years (Katehi, Pearson, & Feder, 2009). In conjunction with engineering design problems and projects a variety of models, methods, rubrics, and guides have been developed to assist with the assessment of these open-ended scenarios Denson, Buelin, Lammi, & D'Amicom 2015; Diefes-Dux, Moore, Zawojewski, Imbrie, & Follman, 2004; Schilling & Applegate, 2012).

Despite a wide-range of possible options for assessment, the highly-creative and open-ended nature of these problems makes reliable, valid, and efficient assessment difficult (Bartholomew, 2017; Kimbell, 2007). The inherently large number of possible solutions to an open-ended problem, coupled with the often highly-creative nature of student designs, has perplexed and troubled those tasked with assessing these projects (Pollitt, 2004, 2012; Pollitt & Crisp, 2004). Rubrics and criterion-based approaches to grading have been lauded as a potentially viable solution for these situations (Kimbell, 2007, 2012a; Denson, Buelin, Lammi, & D'Amicom 2015), however teacher bias and subjectivity have continued to make valid and

reliable assessment difficult with a purely rubric-based approach (Bartholomew, 2017; Pollitt, 2012).

While studying the problems inherent in the assessment of open-ended design problems, Kimbell and Pollitt, two researchers from the United Kingdom, piloted the innovative assessment method of ACJ. ACJ, based on research and work originally done by Thurstone (1927), centers on principles of comparative judgment. Thurstone argued that human comparative judgments (judgments between two items) are more valid and reliable than decisions based solely on a predetermined rubric (Pollitt, 2004; Thurstone, 1927).

In an ACJ assessment setting a teacher, or judge, does not use a rubric to tally a score for each student; rather, the teacher/judge simply views sets of student work and identifies which item is "better" based on a predetermined rubric and their own professional expertise. As the teacher/judge repeats this process—identifying the better in a set of two items—until a rank order develops for all the student work. The ACJ method of assessment can be undertaken by a single teacher/judge or as a group-effort between a group of teachers/judges. While ACJ can be an individual task or one undertaken by multiple judges each additional judge participating in an ACJ session works to reduce the overall number of judgments required by each participating judge to complete the assessment. Other advantages of utilizing multiple judges revolve around the assessment of inter-rater reliability and the ability to draw on opinions, values, and expertise of multiple judges.

As student work is compared, each item attains a "win-loss" record with a "win" rising from their work being picked over another item and a "loss" stemming from another piece of work being picked over their own work. Using a tested and validated process and algorithm (Pollitt, 2012) the ACJ engine can become "adaptive" in the sense that the pairings are increasingly similar (similar win-loss records) and the judgments become more difficult and finely tuned. The result from the ACJ process is a rank-order of student work from the "best" to the "worst." ACJ has been tested and validated repeatedly (Bartholomew, et al., in press; Hartell & Skogh, 2015; Kimbell, 2012a, 2012b; McMahon & Jones, 2015), in a variety of settings and with a plethora of problems, and consistently shown extremely high levels of validity and reliability. While ACJ assessment may not apply to every scenario, it presents uniquely valid and reliable approach for assessing open-ended design problems that have traditionally been very difficult to assess (Kimbell, 2012a). Recent technological advances have made the ACJ judgment process easier to implement through software platforms and a variety of free and paid-for options for ACJ assessment currently exist. Questions around the feasibility and efficiency of ACJ as a tool for assessment need to be addressed and investigated further as the current understanding of these elements of ACJ is incomplete (Bartholomew et al., 2017).

It is important to note that ACJ is not simply a return to norm-referenced grading. Although the final rank order resulting from the ACJ process does represent a norm-referencing of student work the manner in which grades are applied is still subject to the teacher's discretion. Utilizing the rank order to provide a norm-referenced grading approach is only one method of grade distribution and a variety of other grading methods have been employed using the final rank order from the ACJ approach (Bartholomew, 2017).

In addition to the final rank order emerging from the ACJ process, other pieces of valuable information can be collected including a "misfit statistic" for each judge and student work piece, and a comment/rationale for each judgment. In ACJ assessment the *Rasch* model misfit statistic represents how closely each judge's decisions align with the final rank order of all the assessed work and the misfit statistic for each item represents how consistently each item was compared with the other items (see Pollitt, 2012a and 2012b for a full discussion of the *Rasch* statistical methods in ACJ). By using the misfit statistic, one can identify judges whose assessment practices out of line with the group as a whole enabling further training for judges. Misfit statistics for individual pieces of student work can help identify potentially problematic pieces of student work requiring further analysis and assessment. Current versions of ACJ assessment software can also facilitate a comment/rationale for each judgment—an opportunity for judges to justify their decisions for each comparison. These comments can "follow" the respective student items and be used to increase learning and understanding of the final ranking for each item (Bartholomew, 2017, Bartholomew et al., 2017).

Methodology

This study collected data from two sources: 1) the design notebooks and testing results of designs from 16 undergraduate engineering students (4 female and 12 male) who were within their first-year of an engineering major, and 2) the ACJ ranking of student portfolios by a panel of five judges with a background in assessing design. The student participants had an average age of 20 years and were enrolled in the first required introductory engineering course at one research-intensive public university in the Appalachian region of the southern United States. The judges were selected because of their background in education, from middle-school to higher education, and their experience with teaching and assessing design.

Student Design Data. The student design data included engineering notebooks, solution prototypes, and prototype performance results. All students worked on the same open-ended engineering design problem (see Figure 1) which was presented to them immediately following their introduction to the engineering design process in their course. Students were tasked with designing, constructing, and evaluating a system or device to reduce the turbidity of contaminated drinking water in a developing nation after the onset of a natural disaster. Students worked individually in various laboratories and were provided access to a variety of materials, tools, and supplies while engaged in the challenge. Students were also provided with a turbidity sensor, connected to a computer interface, which allowed them to evaluate how well their device removed potential contaminates from a water sample. On average, the participants completed the challenge within one hour, 21 minutes, and 16 seconds. Prior to assessment, all identifiable information was removed and student portfolios were assigned a letter.

Traditional Assessment. While designing a solution to the design challenge, each student utilized an engineering notebook to document information and design ideas—students were informed they could use the notebooks in any way they deemed necessary and were not required to record any specific type or amount of information related to the task. The provided turbidity sensor also recorded student design data as the students progressed through multiple design iterations and tests of their prototype. At the completion of the design challenge all student work was prepared for grading by the course instructor. In order to facilitate grading, all notebooks were collected and combined with pictures of each student's final design and turbidity

data (see Figure 2). The course instructor assessed each student individually in alignment with the design challenge and a standard grading rubric for open-ended problems. Each student was given a score on several sub-component items, related to the design challenge, and these sub-component scores were totaled to form an overall score for each student.

ENGINEERING DESIGN CHALLENGE

Introduction

In many developing countries, clean water is not readily accessible and therefore disease and illness is spread. This is especially true in the aftermath of natural disasters in these areas. While there are many challenges related to clean water, purification is an important part of many water treatment processes.

Problem Statement

People in developing countries do not have continuous access to clean water, especially after the onset of a natural disaster. Water in these situations needs significant purification. However, water purification units are expensive and not easy to obtain. Therefore, you are tasked to design an inexpensive, easy to use, easy to assemble, durable, and low maintenance water purification system using low cost, readily available materials to quickly remove contaminants from water. You will focus on reducing the turbidity of a sample of water.

Testing Performance

Turbidity is a measure of the lack of clarity (cloudiness) of water and is a key test of water quality. Turbidity is apparent when light reflects off of particles in the water. Some sources of turbidity include soil erosion, waste discharge, urban runoff, and algal growth. In addition to creating an unappealing cloudiness in drinking water, turbidity can be a health concern. It can sustain or promote the growth of pathogens in the water distribution system, which can lead to the spread of waterborne diseases. Turbidity is measured in Nephelometric Turbidity Units, NTU. Water is visibly turbid at levels above 5 NTU. However, the standard for drinking water is typically 1.0 NTU or lower.

Prototype Materials

- You are not limited to any specific materials.
- You can use any materials necessary to create the best solution.
- You should not be concerned with material availability.
- You should design your solution to best meet the specified criteria and constraints.

Equipment/Supplies

- Computer and Internet access, distilled water, contaminated water, water sample bottle with lid, paper towel, bucket, Vernier turbidity sensor/equipment, LabQuest Mini, Logger Pro software

Deliverables

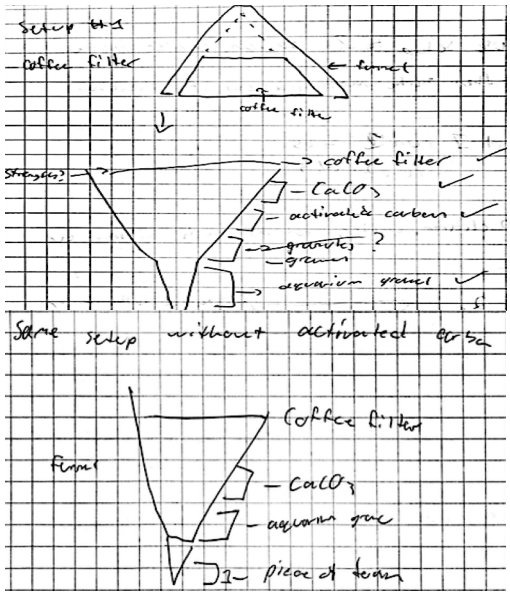
- Functioning Prototype of Quality Construction
- Project Journal
- Solution Analysis - A summary of the details of the design, its benefits, uses, and other important information that explains the design solution.

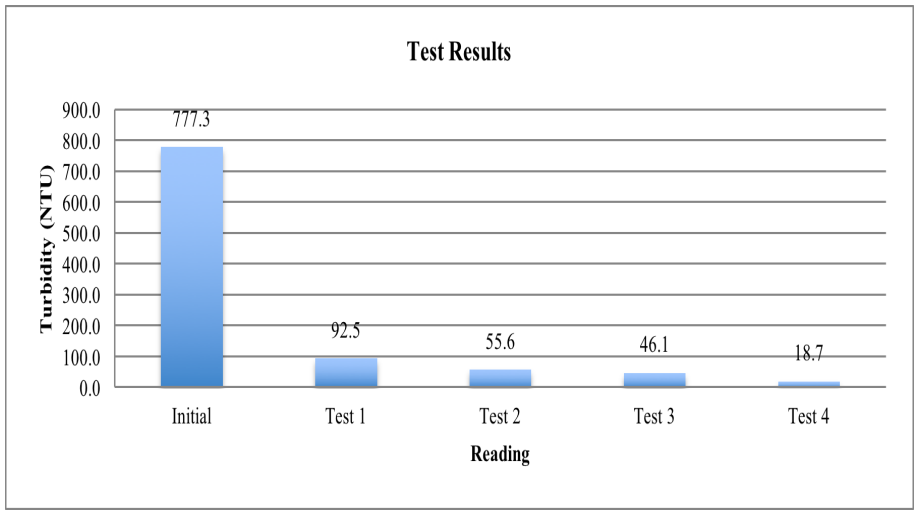
Figure 1. Engineering design challenge.

ACJ Assessment. Following the traditional grading by the instructor, all student work was prepared for the ACJ method. The preparation process involved scanning all pages from the student design notebooks and combining them with a photograph of the student work into a single pdf file for viewing in the ACJ assessment engine (see Figure 3). The student turbidity

results were intentionally excluded from these artifacts in an effort to reduce potential bias in judgment.

Participant Design Summary

Design	**Prototype**
	



Solution Analysis: 1st filter funnel, two piece of foam at the bottom of funnel. Filled funnel with aquarium gravel, gravels, activated carbon, Caco3, then place coffee filter on tap. 2) Poured original sample in, that reading was 423. Water came out black, still black that had lower NTU. 3) Took original resulting water, poured it back it, but it went slower. Enough water came out that was able to be tested. Got a reading of 61.3. 4) Tested activated carbon alone and it made water black as the original resulting sample. So I took out activated carbon and ran Setup 2 with 1 piece of foam, aquarium gravel and CaCo3. Resulting NTU was higher than original NTU. 5) CaCo3 is essential to the filter but should be run several times through (Two times in Setup1). 6) What to do in the future: Make two setups, exactly the same, 1 piece of foam on the bottom, but run water through each setup twice. Take resulting water from Setup1 and run it through Setup 2.

Figure 2. Sample of the design data collected from each participant.

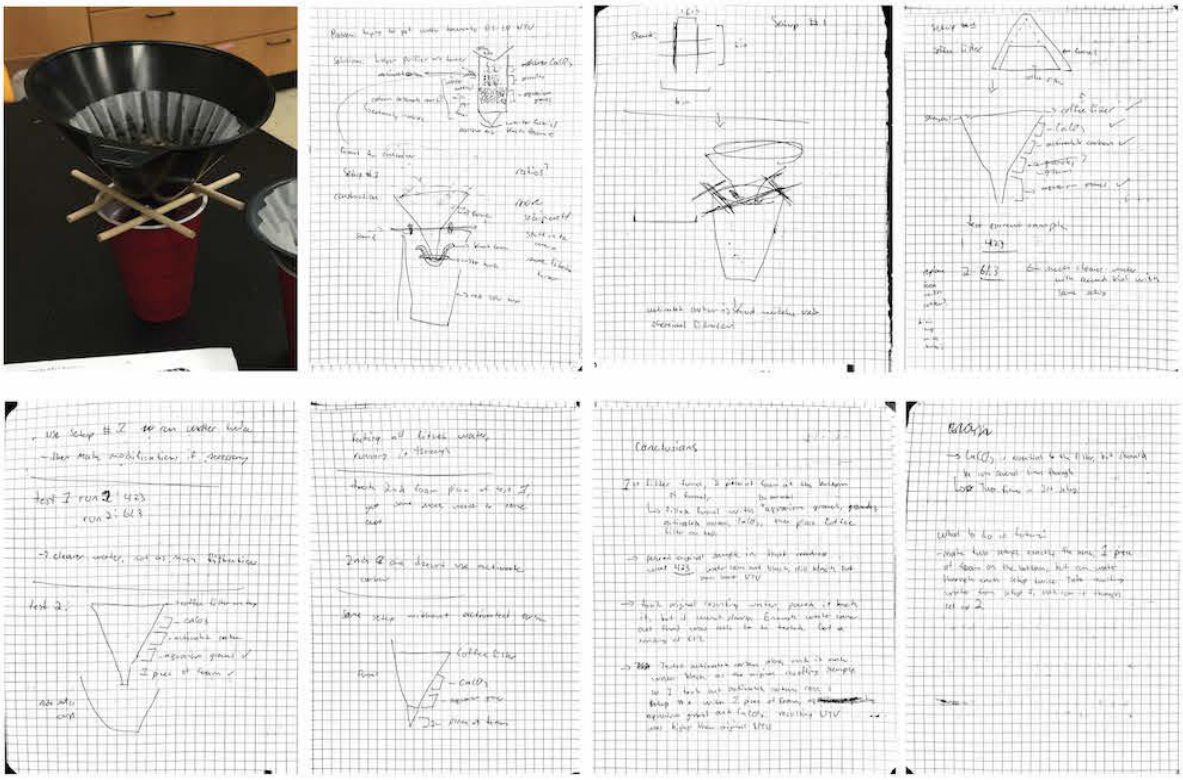


Figure 3. Sample of the student design work presented in the ACJ software. Judges were capable of zooming in to examine the details of the design work.

The panel of five judges, which did not include the course instructor who completed the traditional assessment of the student work, was assigned to judge the student artifacts using the ACJ software *CompareAssess*. Each judge attended a preliminary training session where they were introduced to the student design task, the ACJ process, and the technology-enabled ACJ assessment engine (*CompareAssess*).

Following the training, each judge completed 10 comparative judgments which resulted in a total of 6 *rounds* of judgment—a *round* of judgment in ACJ occurs each time every item is compared at least one time with another item. All judges attended a follow-up meeting to discuss their experience and identify common criteria for comparisons. Using the assignment, the rubric, and their background and experience thus far, the judges reached consensus to continue making judgments with a specific emphasis on: 1) student evidence for justifying design decisions, 2) detailing of design plans, and 3) action based upon design analysis while making the judgments.

Using these common criteria each judge was asked to complete 10 additional comparative judgments. All judges completed at least 10 additional judgments with one judge opting to continue judging through 22 additional comparisons. The resulting rank order of

student work was recorded and a reliability statistic representing the repeatability of the concluded rank order was calculated.

Findings

The findings from this study were derived from a variety of sources and prepared for analysis using SPSS statistical software (Version 23). The specific data, used in an effort to explore each of the research questions, was obtained from:

- student scores from traditional assessment—as assigned by their course instructor,
- student design turbidity scores—as calculated by the turbidity computer interface provided to each student,
- the final rank order of student work—as derived from the ACJ performed by a panel of five judges,

Each of the research questions, and the associated findings, will be discussed in turn.

Research Question 1: What relationship, if any exists, between the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment? At the conclusion of the engineering design challenge, all student work was collected and assessed by the course instructor using a traditional assessment approach. Students were scored on several sub-component categories and then assigned a total overall score using the sum of all sub-component scores (see Table 1).

Table 1
Student Scores from traditional rubric assessment

	Student Portfolios															
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**
Research	2	1	2	3	1	3	2	2	3	3	3	1	2	1	3	2
Multiple Solutions	3	4	2	3	2	2	2	1	3	2	3	2	2	1	4	2
Justification	2	3	4	4	2	3	2	1	3	4	4	3	1	2	4	2
Material Selection	3	3	4	5	3	3	1	2	3	4	4	3	4	3	4	2
Performance	1	5	4	5	2	4	2	3	4	4	4	3	4	2	3	3
Durability	1	1	4	5	1	3	2	2	3	4	4	3	4	3	4	1
Ease	1	2	3	5	2	2	2	2	3	3	3	3	4	2	4	2
Notebook	3	2	4	4	2	3	2	1	3	3	3	2	2	2	4	1
Testing	2	4	4	5	3	3	3	3	4	3	4	4	3	3	5	3
Revision	2	2	3	4	2	2	2	2	4	4	4	3	2	3	5	2
Total Score	**20**	**27**	**34**	**43**	**20**	**28**	**20**	**19**	**33**	**34**	**36**	**27**	**28**	**22**	**40**	**20**

Following the traditional grading of student work, all student portfolios were prepared and uploaded into the ACJ grading engine *CompareAssess* and judged by a panel of five judges. The resulting rank order for student portfolios (see Table 2) was computed following 14 *rounds* of judgment and corresponds with a reliability level of $r = .95$, signifying a high-level of repeatability for results with the study parameters.

Table 3
Student portfolio rank order

Rank	Portfolio
1	Student O
2	Student D
3	Student C
4	Student M
5	Student I
6	Student L
7	Student K
8	Student A
9	Student B
10	Student J
11	Student G
12	Student F
13	Student N
14	Student E
15	Student P
16	Student H

A Spearman correlational analysis was computed using the final rank order for student portfolios and the final score obtained by each student through the traditional assessment from their course instructor. The resulting Spearman's rho value (R_s = -.79) was significant at the p < .01 level. It should be noted that the negative correlation is not surprising as a lower rank order suggests a better portfolio. The high-level of correlation between the final rank order, obtained through ACJ assessment, and the score obtained by students through traditional assessment from their course instructor, suggests alignment between traditional grading approaches and ACJ assessment. Therefore, the ACJ method can produce a rank order that if used for grading purposes aligns to the grades that can result from a traditional rubric.

Research Question 2: What relationship, if any exists, between the performance of student design projects and the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment? The turbidity scores for each student were obtained using a computerized turbidity sensor which recorded student results as they progressed through the design challenge. Student results (see Table 4) were recorded and the lowest turbidity achieved was identified (marked in bold) for each student. It should be noted that two students did not follow the provided protocol – rather than obtaining a new sample of turbid water for each testing these students repeatedly used the same sample water, introducing the possibility for conflated results.

Table 4

Student Water Turbidity Test Results

	Turbidity Level						
	Initial	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**
Student A	780.8	780.8	780.8	**780.8**	-	-	-
Student B	780.8	71	**2.7**	35.7	28.8	28.4	-
Student C	780.8	194.2	108.7	**58.3**	-	-	-
Student D	780.8	155.7	63.2	31.5	16.6	4.6	**0.1**
Student E	767	148.5	**104.3**	-	-	-	-
Student F	767	76.2	22.5	21.5	**19.8**	-	-
Student G*	767	366.1	369.4	140.6	**105.4**	122.3	-
Student H*	767	336	167	**54.1**	-	-	-
Student I	777.3	**24.1**	30.4	54.4	349.5	-	-
Student J	777.3	61.1	**37.9**	-	-	-	-
Student K	777.3	18.7	92.5	55.6	**46.1**	-	-
Student L	777.3	276.5	**66.1**	-	-	-	-
Student M	716.8	360	165.4	33.5	**13.3**	-	-
Student N	716.8	218.5	223.2	**147.4**		-	-
Student O	716.8	423.8	**61.2**	836.7	-	-	-
Student P	716.8	106.6	38.7	**18.4**	-	-	-

Note. Participant test results are reported in Nephelometric Turbidity Units (NTUs). Water is visibly turbid at levels above 5.0 NTUs and the standard for drinking water is typically lower than 1.0 NTU. * Participant did not obtain a new sample of turbid water for each test.

After obtaining the lowest turbidity value for each student a Spearman correlational analysis was calculated to determine the relationship between the performance of each student design—as measured by their lowest turbidity value—and the final rank order and score obtained through traditional grading approaches (Table 4). The test returned no significant correlation between the turbidity results of student designs and either traditional grading or ACJ assessment approaches, suggesting a lack of correlation between the grading of student work and the actual performance of student designs.

Table 4

Spearman rank order correlation table for turbidity, traditional grading score, and ACJ rank (N = 16)

Variables	1	2	3
1. Turbidity score	—		
2. Rubric score total	−.42	—	
3. ACJ rank	.20	−.79**	—

** $p < .01$

Discussion and Implications

This research centered on the design portfolios of 16 undergraduate engineering students tasked with an open-ended engineering design problem which required students to reduce the turbidity levels of water. Student work, including design portfolios and turbidity levels, was

collected and assessed using a traditional approach to grading by their course instructor. Student portfolios were then assessed by a panel of five judges using ACJ. The guiding research questions for this study were:

1. What relationship, if any exists, between the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment?
2. What relationship, if any exists, between the performance of student design projects and the assessment results of an open-ended engineering design problem using traditional rubric and ACJ approaches to assessment?

The relationship between traditional and ACJ approaches to assessment. Our analysis revealed a significant correlation between traditional and ACJ assessment. This was encouraging as it strengthens the validity of ACJ as a form of assessment which will yield results comparable to currently utilized forms of assessment. These findings support other research related to the validity of ACJ assessment (Seery, Canty, & Phelan, 2012) and support calls for further implementation of ACJ into educational settings (Bartholomew, 2017; Bartholomew et al., 2017; Bartholomew & Strimel, 2017; Kimbell, 2012, 2017)**.**

The relationship between student design performance and traditional and ACJ approaches to assessment. Interestingly neither traditional nor ACJ assessment approaches were significantly correlated with performance of the student designs. These findings are similar to other findings which found that student performance on design portfolios does not always align with student product performance (Bartholomew, 2016; Bartholomew et al., 2017). The implications of this finding may be far-reaching—if a student's grade on an assignment does not reflect the actual quality of the final product, is the assessment approach effective? These considerations are important as teachers work to form assessments which accurately reflect the quality of a student's work.

Conclusion

This study found that ACJ and traditional forms of assessment were significantly correlated. This suggests that the results from ACJ assessment techniques align closely with traditional approaches to grading student work and strengthens the validity of the ACJ approach to assessment. When compared, no significant relationship was found between either ACJ or traditional approaches to assessment and the actual performance of student designs. This is important as most grading and assessment techniques are assumed to be a measurement of student success. There is need for additional research, study, and conversation around defining and measuring success in open-ended design scenarios.

The findings from this research are limited in scope as ACJ is compared with only one form of assessment utilized by one instructor for a specific design problem. Further research with a variety of traditional assessment approaches, rubrics, students, and design problems would

assist in clarifying the relationship between ACJ and traditional grading approaches. Further research is needed into the implications of ACJ assessment and its impact on students, teachers, and judges. Research is also needed related to how judges make decisions in ACJ settings and how these decisions correlate with the decision-making process of teachers utilizing traditional grading approaches.

Importantly, neither assessment in this study was significantly correlated with the actual results of student work. While limited to the study parameters, this finding suggests a need for further research into the correlation between assessment results and actual performance of student designs. New or innovative approaches to the assessment of student designs may be needed in order to accurately reflect the success of a student's work.

ACJ appears to be a valid, reliable, and plausible approach to assessment and aligns well with current assessment strategies in place. Additional research into ACJ, and its' place in educational assessment settings, is needed. Discussion and research into the relationship between assessment results and actual design results is necessary to clarify a potential disconnect which may be adversely impacting students.

References

Bartholomew, S. R. (2016). *A Mixed-Method Study of Mobile Devices and Student Self-Directed Learning and Achievement During a Middle School STEM Activity* (Doctoral dissertation, Utah State University).

Bartholomew, S. R. (2017). Assessing open-ended design problems, *Technology and Engineering Education Teacher* (In Press)

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* (In Press).

Bartholomew, S. R. & Strimel, G. (2017). The Problem with assessing open-ended problems. *Techniques* (In Press).

Cunningham, C. M. (2009). Engineering is elementary. *The Bridge, 30*(3), 11-17.

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), pp. 23-40.

Diefes-Dux, H. A., Moore, T., Zawojewski, J., Imbrie, P. K., & Follman, D. (2004, October). A framework for posing open-ended engineering problems: Model-eliciting activities. In Frontiers in Education, 2004. FIE 2004. 34th Annual (pp. F1A-3). IEEE.

Dutson, A. J., Todd, R. H., Magleby, S. P., & Sorensen, C. D. (1997). A review of literature on teaching Engineering design through project-oriented capstone courses. *Journal of Engineering Education, 86*(1), 17-28.

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).

Goel, V., & Pirolli, P. (1992). The structure of design problem spaces. *Cognitive science, 16*(3), 395-429. doi: 10.1207/s15516709cog1603_3

Jonassen, D. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development, 48*(4), 63-85. doi: 10.1007/BF02300500

Katehi, L., Pearson, G., & Feder, M. (2009). Engineering in K-12 education. Committee on K-12 Engineering Education, National Academy of Engineering and National Research Council of the National Academies.

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

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

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

Kimbell, R. (2017). *Making assessment judgements: policy, practice, and research*. Manuscript in preparation.

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

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, September). Could comparative judgments of script quality replace traditional marking and improve the validity of exam questions? Paper presented at the BERA Annual Conference, UMIST Manchester, England.

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

Pollitt, A. (2012b). Comparative judgment for assessment. *International Journal of Technology and Design Education, 22*(2), 157-170.

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.

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. http://doi.org/10.3163/1536-5050.100.4.007

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

Whitehouse, C., & Pollitt, A. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. Manchester: AQA Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/sites/default/files/pdf_upload/ CERP_RP_CW_20062012_2.pdf.