Chapter 1

A systematic review of research around Adaptive Comparative Judgment (ACJ) in K-16 education

SCOTT R. BARTHOLOMEW
Dept. of Technology Leadership and Innovation, *Purdue University, West Lafayette, Indiana 47907, USA*

EMILY YOSHIKAWA-RUESCH
Dept. of Technology Leadership and Innovation, *Purdue University, West Lafayette, Indiana 47907, USA*

Abstract

While research into the effectiveness of open-ended problems has made strides in recent years, less has been done around the assessment of these problems. The large number of potentially-correct answers makes this assessment difficult. Adaptive Comparative Judgment (ACJ), an approach based on assessors/judges working through a series of paired comparisons and selecting the better of two items, has demonstrated high levels of reliability and effectiveness with these problems. Research into using ACJ, both formative and summative, has been conducted at all grade levels within K-16 education (ages 5-18), with a myriad of findings. This paper outlines a systematic review process used to identify articles and synthesizes the findings from the included research around ACJ in K-16 education settings. The intent of this systematic review is to inform decision-makers weighing the potential for ACJ integration in educational settings with researched-based findings around ACJ in K-16 educational settings. Further, this review will also uncover potential areas for future researchers to investigate further into ACJ and its' implications in educational settings.

Key Words: Adaptive Comparative Judgment, Open-ended problems, Assessment

Introduction

The preparation of students for future employment and an emphasis on Science, Technology, Engineering, and Mathematics (STEM) education and skills has led to a larger emphasis on the integration of open-ended problems in education (Bartholomew, 2017; Dearing & Daugherty, 2004; Diefus-Dux, et al., 2004; ITEEA, 2000/2004/2007; NAE & NRC, 2014; NRC, 2009; Reeve, 2015; Sanders, 2009; Wicklein, 2006). This emphasis, often joined with problem- and project-based learning, has aimed at better preparing students for success in highly flexible and technologically-driven work environments (Dearing & Daughterty, 2004). Despite widespread efforts around open-ended problems, and their integration in education, much less has been done around the assessment strategies and techniques associated with these types of problems (Bartholomew, 2017; Kimbell, 2007, 2012a, 2012b; Pollitt, 2004, 2012; Pollitt & Crisp, 2004). Open-ended problems, with a myriad of potentially correct solutions, have traditionally been very difficult to assess with validity, reliability, and efficiency (Bartholomew, 2017a, 2017b).

Although rubrics, portfolios, technology-enabled platforms, and criterion-grading tools have all been employed towards improving the assessment of open-ended problem many of the challenges (e.g., reliability, efficiency) have remained. Research and experience continue to affirm that a teacher's ability to assess open-ended problems with fidelity using traditional forms of assessment is poor at best—past experiences, personal preferences, time in the profession, and a variety of other factors all "muddy the waters" and contribute to difficulty in assigning grades reliably (Alkharusi, 2011; Bartholomew, 2017; Crossman, 2004; Dietrich, 2010; Kimbell, 2007, 2012a, 2012b, 2016; McMillan & Nash, 2000; Pollitt, 2004; Rice, 2010; Westerman, 1991). However, a recently revisited approach to assessment titled Adaptive Comparative Judgment (ACJ) has been increasingly utilized in recent years with success in addressing many of the challenges associated with open-ended problems (Bartholomew, 2016; Bartholomew, Strimel, & Jackson, 2017; Hartell & Skogh, 2015; Kimbell, 2007, 2012a, 2012b; Kimbell, et al., 2007; Newhouse, 2011; Seery, Canty, & Phelan, 2012; Steedle & Ferrara, 2016).

ACJ was originally conceptualized as Comparative Judgment (CJ) in the 1920s by psychologist Louis Thurstone (1927) who presented several alternative methods of constructing measurement scales for assessment. Comparative Judgment is a process where a judge/assessor views two items and chooses the better of the two items. This process assumes that as judges/assessors view items they assign an instinctive value to each item based on their expertise, past experiences, and the item's quality. Thurstone posited that when two phenomena are placed in comparison with one another, an individual is able to use their own instinctively-assigned values for each item to compare and identify which of the two phenomena are 'better' with great levels of fidelity. Thurstone demonstrated that by repeatedly comparing pairs of items a rank-order could be produced of all the items assessed with very high levels of reliability. This approach to assessment, which demonstrated highly-reliably results, was largely unused for decades—largely as a bi-product of the arduous time-requirements associated with the repetitive comparison process.

Decades later, Thurstone's work was revisited by Pollitt and Murray (1996) who saw the opportunity to utilize technology as a means of optimizing this process. Pollitt and Murray (1996) used Thurstone's ideas, in conjunction with Georg Rasch's mathematical models for educational tests (Rasch, 1993), to further develop the idea of comparative judgment as a tool for assessment. Initial piloting of this approach demonstrated markedly more reliable results than traditional approaches to assessment (Kimbell, 2007; Pollitt & Whitehouse, 2012), especially in relation to open-ended problem assessment (Kumar & Natarajan, 2007).

In addition to the use of technology for facilitating the comparative judgment process, work was done to develop an algorithm which adaptively paired similarly-ranked items and worked to further reduce the time required for completing the assessment process (Kimbell, 2012a, 2012b; Pollitt, 2012). The addition of the algorithm—which adaptively pairs similarly-ranked items for assessment—to the process led to the concept of *adaptive* comparative judgment (ACJ). With the applied algorithm, improved reliability can potentially be achieved after fewer comparisons than traditional comparative judgment which relies on random pairings (Bramley, 2015; Steedle & Ferrara, 2016).

This approach to assessment, although markedly different from other assessment techniques, has been implemented by a variety of individuals in different locations, subject areas, and with different age groups (Bartholomew, 2017a). However, not all reviews have been positive with Bramley (2015) offering the harshest critique of the approach. Bramley (2015) challenged the reliability of the rank-order produced through ACJ explaining that the adaptive aspect of ACJ inflates the reported reliability. However, Pollitt (2015) countered Bramley's arguments, explaining that the demonstrated inflation is trivial and that "errors that appear only in the third decimal place of an alpha coefficient are of no practical importance at all" (p. 8). Continued efforts towards investigating the reliability, validity, and feasibility of ACJ, as an approach to assessment in open-ended problems, are ongoing in a variety of settings with the dominant technology tool for implementing ACJ being marketed by *DigitalAssess* as internet browser-based platform titled *CompareAssess* (DigitalAssess, 2017).

Although ACJ appears to be gaining traction in educational settings, the body of research related to ACJ in these settings has not been synthesized. Therefore, we sought to perform a systematic review of literature related to ACJ in K-16 settings which may serve as a starting point for understanding this process, its' implementation into educational settings, and the potential benefits and challenges of utilizing this approach. We intend this piece to be a useful tool, with research-based conclusions, for decision-makers weighing the potential for ACJ's implementation in educational settings. The guiding question for this review was:

> What are the key findings related to research around the implementation of Adaptive Comparative Judgment in K-16 education settings?

Method

Systematic Literature Reviews

Consistent with our intent, the guiding research question, and recommendations of Borrego, Foster, and Froyd (2014) around systematic reviews of literature, we investigated the current literature around ACJ in K-16 settings. This effort involved collecting studies conducted on the topic, refining and narrowing the results, and highlighting key findings related to the research question. This work is not intended to include every item of work related to ACJ; rather, this work is intended to serve as a starting point for individuals interested in ACJ and its' implementation in K-16 education (ages 5-18). Indeed, as Petticrew and Roberts (2008) suggest in relation to systematic literature reviews, we aim to provide a "general overall picture of the evidence in a topic area" (p. 21). As such, this work will highlight articles, related to ACJ and its' implementation, which work to inform our guiding research question and may serve useful in future efforts around ACJ in K-16 settings.

Search Parameters

To begin the process of identifying relevant literature and related search parameters several prominent articles, centered on ACJ for assessment, were reviewed. Prominent articles were selected based on their citation in numerous (>5) ACJ-related publications. These publications and the accompanying cited works were used to establish initial search parameters for investigation. Following the review of these articles an additional search was conducted using the key words "adaptive comparative judgment" and "ACJ" in academic journal search engines related to education (e.g., GoogleScholar, ERIC, EBSCOhost, and Education Full Text). Additionally, as ACJ is highly-connected, and often confused with "comparative judgment," both "comparative judgment" and "CJ" were also used in the search engine efforts.

These efforts yielded 133 total results on "Education Source," with "ERIC" producing 97, "Education Full Text" providing 65, and Google Scholar producing about 1,400,000 results. Review of these results showed that the vast majority of these articles were not relevant to adaptive comparative judgment or its' implementation in education settings. Further focusing our search results we constrained the search to "Adaptive Comparative Judgment" as a key search phrase. Utilizing this as the key search phrase, "Education Source" returned eight results (with two of the results referring to the same article), "ERIC" produced four results, "Education Full Text" yielded two results, and "Google Scholar" produced 40 results. After removing duplicates 46 total articles were collected; all of these articles were published after 2012.

In addition to the literature search, contact was made with leading researchers and publishers of research related to ACJ. Through these efforts an additional 35 sources (e.g., items such as conference papers and/or unpublished works) were added resulting in a total of 81 articles and papers for further review and analysis. One of these papers was removed because it was not written in English which resulted in a total of 80 papers for review.

The next step in the process involved classifying, and then removing, items based on several predetermined criteria. This was done by reviewing the abstracts, introduction, methods, and findings sections for each work. Articles that did not show to meet the criteria were removed. The classification categories and the criteria for inclusion are listed here:

1. **Seminal papers:** highly-cited papers around ACJ which were influential in the development or implementation of ACJ approach to assessment.

   Not Included: papers focused solely on comparative judgment (comparative judgment is similar to ACJ but does not involve the use of the adaptive algorithm to selectively pair items for judgment), rather than *adaptive* comparative judgment. Papers focused on development of "escape" (a large-scale project in the UK which provided the first widespread implementation of ACJ at the outset). Papers focused on the development of the software platform for ACJ assessment rather than research around using ACJ. Papers solely focused on the process of training assessors in an ACJ setting.

2. **Context**: the context of the paper should be limited to studies around ACJ in education settings (K-16).

   Not Included: papers with research founded in settings outside of education or postgraduate studies such as medical school, graduate school, or uses within the workforce. Papers related only to the feedback of judges in ACJ settings or papers dealing solely with self-efficacy and student confidence.

3. **Original Research:** included works should reflect original findings and research around ACJ.

   Not Included: papers from the same author without new findings or explanations.

Following the systematic removal process (see Figure 1) a total of 31 papers remained for further analysis. The majority of the articles which were removed were eliminated in response to criteria 2: the context of the paper should be limited to studies around ACJ in K-16 education settings (K-16). Additionally, many of the papers, which were removed, revolved around *comparative judgment*, rather than *adaptive comparative judgment*.

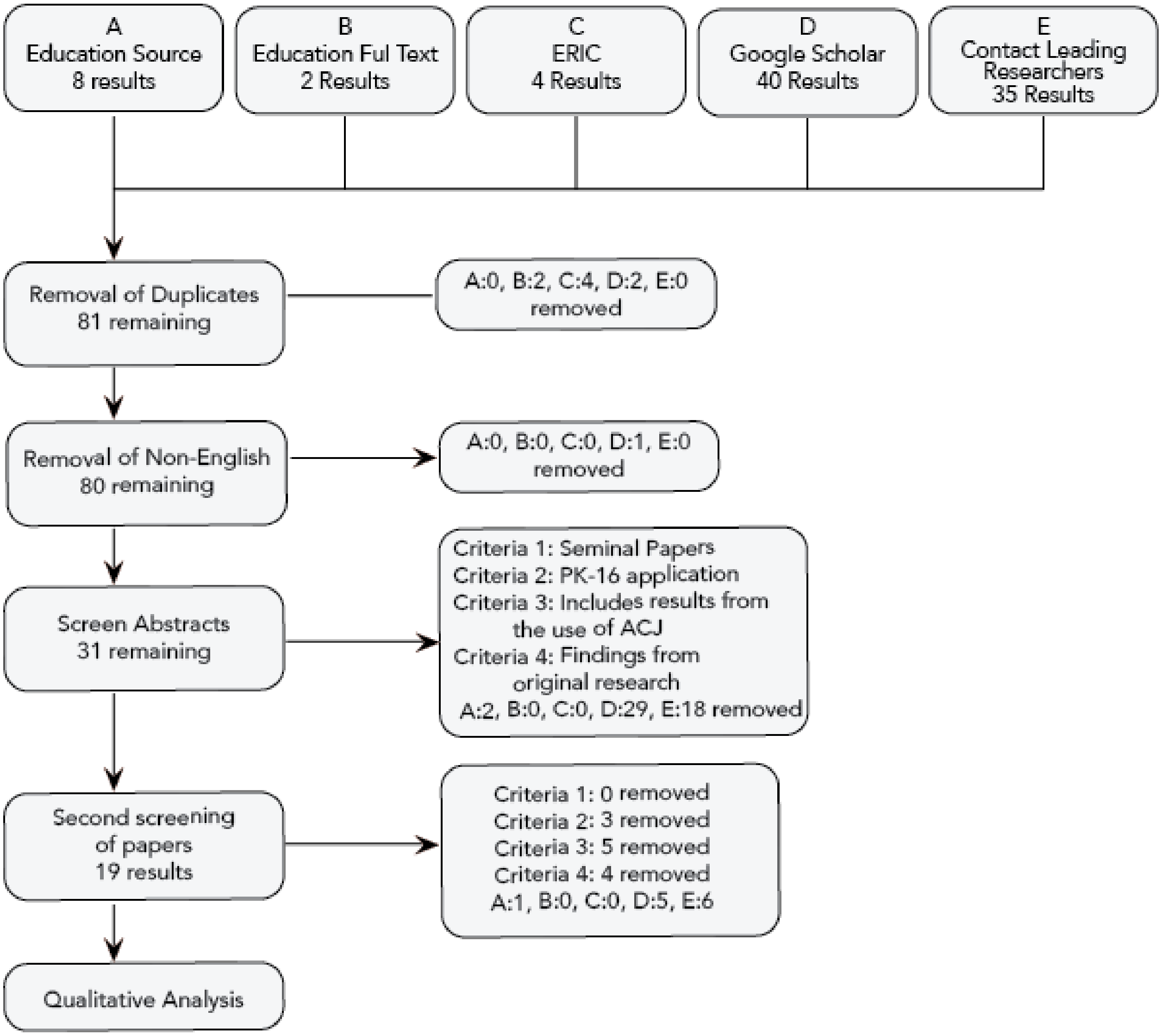

Figure 1. Overview of Systematic Literature Review

Following the screening of abstracts there remained 31 papers that were separated and analyzed again at a deeper level to ensure the criteria were met. This further review revealed that while the abstracts of these papers showed potential for the meeting of the criteria, many of these were duplicates or missing key criteria for inclusion. This second-level sifting resulted in 12 additional papers, out of the 31, being removed for a new total of 19 articles (see Table 1). These 19 articles will serve as the guiding literature for this review.

Table 1

Final Articles for Inclusion in the Synthesis of ACJ-related Literature at the K-16 Education Level

ID**	Title	Author(s)	Source	Year
A	Let's stop marking exams*	Pollit, A.	IAEA Conference, Philadelphia	2004
B	Investigating a judgmental rank-ordering method for maintaining standards in UK examinations*	Black, B., & Bramley, T.	Research Papers in Education, 23(3), 357-373.	2008
C	The validity and value of peer assessment using adaptive comparative judgment in design driven practical education	Seery, N., Canty, D., & Phelan, P.	International Journal of Technology and Design Education, 22(2), 205-226.	2011
D	Summative Peer Assessment of Undergraduate Calculus using Adaptive Comparative Judgement	Jones, I., & Alcock, L.	Mapping university mathematics assessment practices, 63-74.	2012
E	Evolving project e-scape for national assessment*	Kimbell, R.	International Journal of Technology & Design Education, 22, 135-155.	2012
F	The origins and underpinning principles of e-scape*	Kimbell, R.	International Journal of Technology & Design Education, 22, 123-134.	2012
G	The method of adaptive comparative judgment*	Pollitt, A.	Assessment in Education: principles, policy & practice, 19(3), 281-300.	2012
H	Using adaptive comparative judgment to obtain a highly reliable rank order in summative assessment	Pollitt, A., & Whitehouse, C.	AQA: Center for Education Research & Policy.	2012
I	Using digital representations of practical production work for summative assessments	Newhouse, C. P.	Assessment in Education: principles, policy & practice, 21(2), 205-220.	2014
J	Investigating the reliability of Adaptive comparative judgment*	Bramley, T.	Cambridge Assessment, Cambridge, 36.	2015
K	On 'Reliability' Bias*	Pollitt, A	Cambridge Exam Research Technical Report	2015

(continued)

ID**	Title	Author(s)	Source	Year
L	Evaluating comparative judgement as an approach to essay scoring	Steedle, J. T., & Ferrara, S.	Applied Measurement in Education, *29* (3), 211-223.	2016
M	Validity of comparative judgment to assess academic writing: examining implications of holistic character and building a shared consensus	van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S.	Assessment in Education: Principles, Policy & Practice, 1-16.	2016
N	Relationships between access to mobile devices, student self-directed learning, and achievement	Bartholomew, S.R., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., & Nadelson, L.	Journal of Technology Education, *29* (1), 2-24	2017
O	ACJ: A Tool for International Assessment Collaboration	Bartholomew, S.R., Hartell, E., & Strimel, G.	PATT34 Millersville University, Pennsylvania, USA 10–14 July, 2017.	2017
P	A Comparison of Traditional and Adaptive Comparative Judgment Assessment Techniques for Freshman Engineering Design Projects	Bartholomew, S.R., Strimel, G.J., & Jackson A.	International Journal of Engineering Education, *34* (1), 20-33	2017
Q	Illustrating Educational Development Through Ipsative Performance in Design Based Education	Seery, N., Delahunty, T., Canty, D., & Buckley, J.	PATT Conference, Philadelphia.	2017
R	Using Adaptive Comparative Judgment for Student Formative Feedback and Learning During a Middle School Open-ended Design Challenge	Bartholomew, S.R., Strimel, G., & Yoshikawa, E.	International Journal of Technology & Design Education, https://doi.org/10.1007/s10798-018-9442-7	2018
S	Examining the Potential of Adaptive Comparative Judgment for Elementary STEM Design	Bartholomew, S.R., Strimel, G., & Zhang, L.	Manuscript submitted for publication	2018

Note: * denotes a paper determined "seminal" by the authors
** identifiers to be used through the duration of the paper

Mapping Results

The results of our systematic review revealed several important findings related to the state of ACJ-related research in K-16 education. These key areas of synthesis include: context, approach, assessor characteristics, and results; each of these areas will be presented here.

Context. The first step in mapping our results was to identify basic information around the identified research articles. Mapping the results across time (see Figure 2) and by grade-level (see Figure 3) helps to establish the background for the state of current research around ACJ in K-16 education settings. Since 2011, 17 studies around ACJ in K-16 education have been conducted which fit the identified criteria for this systematic review (see Figure 2). There appears to be an upward trend in number of research efforts related to ACJ in K-16 education settings but this should be taken with caution as there are relatively few years for comparison.

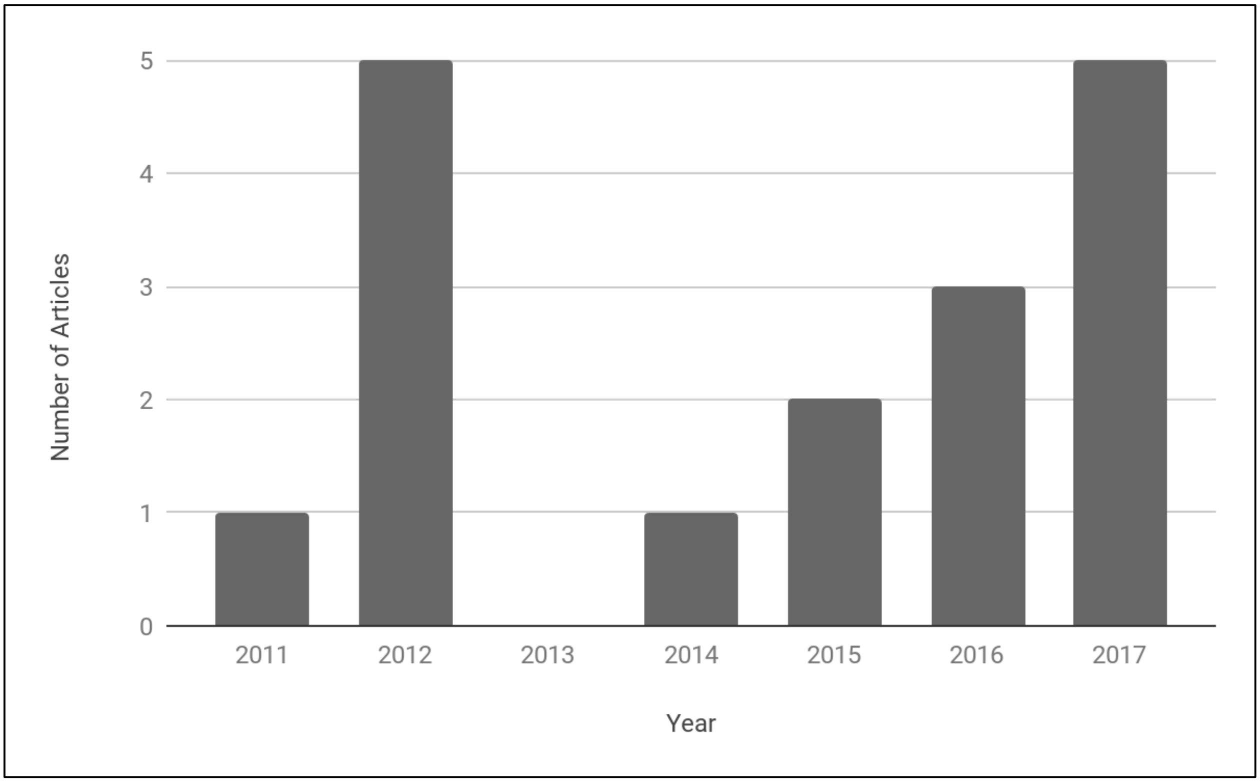

Figure 2. ACJ in K-16 Research Articles by Year

In terms of grade-level research efforts around ACJ our results revealed one study at the elementary level, three at the middle-school level, three at the high school level, and five in the context of undergraduate education (see Figure 3). The majority of early work around K-16 ACJ implementation was conducted at the University level with more recent efforts around the middle school and high school levels (Bartholomew, 2017).

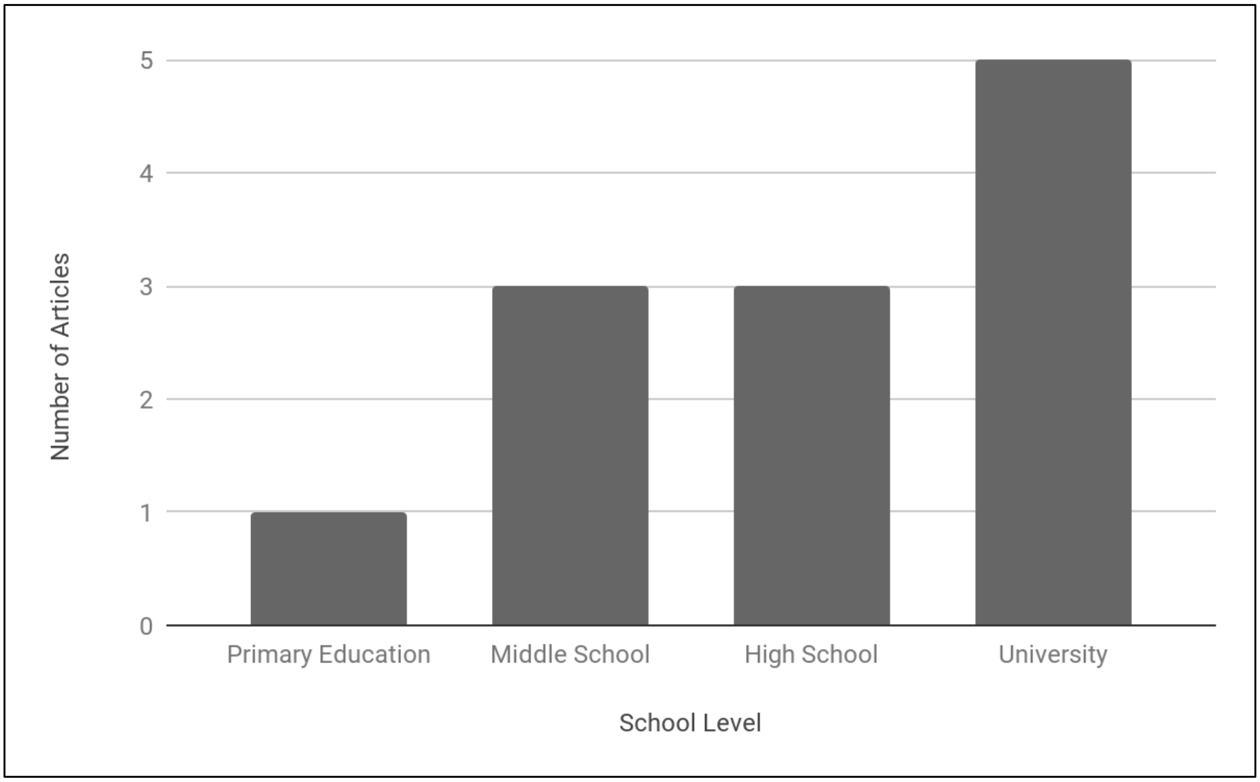

Figure 3. Grade Level of ACJ-Related Research

Beginning in 2004 with the presentation by Pollitt, Elliot, and Ahmed the majority of implementation of ACJ in K-16 settings has continued to employ ACJ in summative settings (D, I, L, P, & Q). These studies have not been confined to one content area but have included English writing (N), design and technology (C, I, N, O, P, Q, & S), human development (N), math (N), social studies (H & P), and teacher-preparation programs (D & I).

Table 2

K-16 ACJ Integration: Settings, Assessors, and Participants

ID	Author (year)	Participants [*n*] & Artifacts	Assessor demographics	Research location	Subject area
B	Black, & Bramley (2008)	University students [10] Compulsory scripts	Assessors	United Kingdom	Psychology
C	Seery, Canty, & Phelan (2011)	University students [137 participating with 63 as judges] Student portfolio	University teachers	Limerick, Ireland	Engineering and Technology Teacher Education
D	Jones, & Alcock (2012)	University students [168] Math scripts	College students	Loughborough University	Calculus
H	Pollitt, & Whitehouse (2012)	High school students [564] Writing Essays	Teachers and Examiners	Greater London Area	Physical and Human Geography
I	Newhouse (2014)	Year-12 students [75 visual arts & 82 design] Design & visual arts projects	External assessors	Western Australia	Design and Visual Arts
L	Steedle, & Ferrara (2016)	High school students [200] Writing responses	Secondary school English teachers	Florida, United States	English
M	van Daal, et al. (2016)	University Students [41] Writing essays	Academic audience, professors, and researchers	Belgium	Academic writing
N	Bartholomew, S.R., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., & Nadelson, L. (2017)	706 middle school students (age 12-13 years old) 200 design portfolios 200 design products	Educators with experience in technology, engineering, and design education	Western United States	Technology & Engineering Education

(continued)

ID	Author (year)	Participants [*n*] & Products	Assessor demographics	Research location	Subject area
O	Bartholomew, S.R., Hartell, E., & Strimel, G. J. (2017)	200 middle school design portfolios and 200 design products (age 12-13 years old)	Educators with experience in technology, engineering, and design education from the United States, Sweden, the U.K., and Ireland	U.S.A., Sweden, U.K., Ireland	Technology & Engineering Education, Design & Technology Education, Teknik
P	Bartholomew, Strimel & Jackson (2017)	University students [16] (average age of 20) Engineering design portfolios	Educators with experience in technology, engineering, and design education	University in Appalachian region	Engineering
Q	Seery, Delahunty, Canty, & Buckley (2017)	University students [128] (Year 3) 128 design tasks	University students	Limerick, Ireland	Initial Technology Teacher Education
R	Bartholomew, Strimel, & Yoshikawa, (2018)	Middle school students [65] (12-13 years-old) Middle school teacher [1] Graphic design projects	Middle school students	Midwestern United States	History
S	Bartholomew, Strimel, & Zhang (2018)	Primary students [92] (age 5-10 years old) Kindergarten teachers [2] Fourth grade teachers [2]	Primary school teachers	Midwestern United States	Elementary school STEM

Approach. Thurstone (1927), who is credited with the original concepts behind ACJ was silent regarding the timing for implementation of ACJ as an assessment tool. The majority of early efforts in ACJ integration centered on summative assessment—mainly for end of year assessment by awarding bodies (M. Wingfield, personal communication, August 29, 2017). Pollitt, Elliott, and Ahmed presented their plans at the University of Cambridge Local Examinations Syndicate (2004) where ACJ was posited to be a replacement for the end-of-year rubric-based approach then employed by the major exam bodies in charge of overseeing the high school examinations.

In addition to several studies around summative assessment through ACJ, recent efforts in utilizing ACJ for formative assessment have also been undertaken. These efforts have largely been confined to design settings with students using ACJ, as a tool for assessment, in the midst of a design project (formative) and then again at the conclusion (summative).

ACJ for Summative Assessment. The majority of K-16 ACJ integration has been aimed towards utilizing ACJ as a tool for summative assessment of open-ended projects. Specifically, the majority of these studies using ACJ for summative assessment (C, O, P, Q, & R) have been conducted in Technology, Engineering, and/or Design classrooms. A brief synopsis of each of the studies in Technology, Engineering and/or Design classrooms is included here:

> **C.** Seery, Canty, & Phelan (2012) implemented ACJ with University students studying Engineering and Technology Education engaged in design scenarios and found that using ACJ for peer judgment allowed students to demonstrate their ability to make critical judgment. They also recognized that by having peer evaluations, the assessor was more fully able to empathize with the work having just gone through the design process. Seery, Canty, & Phelan also reported that "Student confidence in a democratic approach to the assessment formed the basis for unrestricted engagement. As a result, the relationship between student and assessor (their peer) was relaxed" (p. 224).
>
> **O.** Bartholomew, Strimel, & Hartell (2017) employed ACJ for the summative assessment of student work by panels of judges from four different countries. They found high reliability levels for each group of judges with significant correlations across panels from different locations. Their findings suggest ACJ may be a uniquely situated tool for international collaboration in/through summative assessment.
>
> **P.** Bartholomew, Strimel, & Jackson (in press), who used ACJ for summative assessment of college freshman engaged in open-ended engineering problems, demonstrated high levels of reliability and validity by comparing the traditional markings from the instructor with the resulting rank order from the ACJ results. However, they found no significant correlation between the ACJ results and the actual effectiveness of the student designs.

Other subject areas with summative ACJ-related research have included English (L& M), human development (B), math (D), and social studies (H & R). A brief synopsis of each of these studies is included here:

> **D.** Jones and Alcock (2012) employed ACJ in a University level calculus class. In this study, the students participated in peer feedback without given criteria. They looked at

the inter-rater reliability between the peers compared to experts and novices. They found that expert to peer had a 0.63 correlation coefficient, expert to novice had a coefficient of 0.55, and peer to novice had a coefficient of 0.67. Through the process of peer evaluation, students reported recognition of needed self-improvement.

H. Pollitt and Whitehouse (2012) implemented ACJ in high school human and physical geography classes. ACJ was used on 564 high school written essays and the results indicated that ACJ was a valid assessment when compared to traditional marking methods.

L. Steedle and Ferrara (2016) studied the use of ACJ as an assessment for High School English classes from two separate writing prompts. This study found that the total time it may take to use ACJ is greater than traditional methods to reach a suitable reliability level.

M. van Daal, et al. (2016) studied the use of ACJ as a summative assessment for academic writing at a University level. They reported finding ACJ as a valid assessment when grading writing assignments holistically in their study.

R. Bartholomew, Strimel, & Yoshikawa (2017) utilized ACJ for summative assessment with an open-ended design challenge involving middle-school students and found the resulting rank order to be both reliable and valid.

ACJ for Formative Assessment and Learning. In addition to summative assessment, efforts towards using ACJ as a learning tool for students in formative settings have shown great promise in terms of student learning, achievement, and peer-feedback (P). Students who engaged in ACJ formatively have reported increased recognition of areas for improvement, benefits from exposure to peer work, and increased ability to improve their work (P). Efforts in the area of K-16 ACJ for formative assessment include:

Q. Seery, et al. (2017) found that learning gains associated with ACJ were not confined to individual students; rather, they posit that the entire class of students involved in ACJ, for formative learning, will improve significantly over time. Their exploratory data and research has revealed very promising findings to this effect.

R. Bartholomew, Strimel, & Yoshikawa (2017) conducted research with middle school students (age 12-14) engaged in an open-ended design problem. These students utilized ACJ for peer-formative feedback and assessment at the midpoint of their design experience. When compared with the control group of students, the students engaged in ACJ for formative learning performed significantly better than their peers at the conclusion of the assignment ($t(100) = -4.28$, $p < .001$).

Assessors: Training, Selection, and Experiences. In the ACJ process the items being compared are uploaded to a web-based portal which facilitates the pairwise comparisons. In many of the initial implementation settings ACJ was performed by professional assessors or experts (A, E, F, G, H, & K). However, in recent year's movement towards ACJ assessment by a variety of individuals including teachers, industry partners, and students, have all been employed (L, Q, R, & S). Although all the included articles did not provide background on the chosen assessors the information that was provided is synthesized here:

> **B, I, & L.** Three studies (Black & Bramley, 2008; Newhouse, 2014; Steedle & Ferrara, 2016) all used educators or professional assessors for looking at student work.
>
> **C, D, &R.** Three other studies (Bartholomew, Strimel, & Yoshikawa, 2017; Jones & Alcock, 2012; Seery, Canty, & Phelan, 2011) used the students in the participating class as the judges participating in the ACJ sessions.
>
> **M.** van Daal, Lesterhuis, and Coertjens (2016), who looked at the assessment of academic writing, utilized professionals with varied expertise in the writing department for their assessors. Each of these individuals had background and experience in assessment writing products.
>
> **O.** Bartholomew, Hartell, and Strimel (2017) used judges with design, technology, and engineering education backgrounds from several different countries to compare the assessment similarities and differences across location. While all the judges have design, technology, and/or engineering backgrounds the USA-based assessors all held K-12 teaching licenses. The UK-based judges were composed of those with K-12 and college-level backgrounds and the Sweden-based judges were teachers, professors, and graduate students in technology fields.
>
> **P.** Bartholomew, Strimel, and Jackson (in press) compared engineering design portfolios with assessors who all came from backgrounds in teaching the engineering design process. These included practicing teachers, researchers, and graduate students.

In addition to the training, background, and experience of judges several of the research studies have recorded the time required for judges to make the necessary comparisons for the ACJ process. These times were often recorded to compare with traditional grading approaches and the times required have varied greatly between studies (see Table 3), locations, judge background, and subject-area.

Table 3

Judgment Statistics for Assessors

ID	Author	Judge Background	Time (traditional grading)	Time (ACJ)
I	Newhouse (2014)	External assessors	6.4 min for Design 9.9 min for Visual Arts	5.6 min. for Design 5.4 min. Visual Arts
N	Steedle, & Ferrara (2016)	Secondary English teachers	Prompt 1 *M* = 121.2 min. Prompt 2 *M* =116.4 min.	Prompt 1 *M* = 116.7 min. Prompt 2 *M* = 70.5 min.
P	Bartholomew, Strimel, & Jackson (2017)	Educators with experience in technology, engineering, and design education		Average time per judgment ranged from 0:55 to 3:26
S	Bartholomew, Strimel, & Zhang (2018)	Primary Teachers	Project 1 *M* = 6.48 min. Project 2 *M* = 6.87 min.	Project 1 *M* = 11.92 min. Project 2 *M* = 9.80 min.

ACJ Implementation Results. The results from ACJ implementation in K-16 settings have revolved largely around reliability and validity. A detailed discussion of the research around, and statistical approaches to, the ACJ reliability is beyond the scope of this work (see Pollitt, 2012 for a more detailed discussion on these ideas), but several key findings from the included research articles will be presented here related to efficiency, validity, and reliability.

Efficiency. The research included here seems to suggest that ACJ is not a more efficient method of assessment when compared with traditional approaches. However, efforts towards improving the algorithm, and the associated efficiency, have been underway for the past several years by companies such as *Digital Assess* (DigitalAssess, 2015), and some research has suggested that the time required for ACJ is not significantly different from that of traditional assessments (H, L, & S). It should be noted that the time recordings, derived from the ACJ-platform around judge decisions, may not be completely accurate as these times reflect the total time a paired comparison is present before a judge—regardless of how actively focused the judge may be on making the judgment.

Speaking about efficiency, and the ACJ algorithm that guides the process, Steedle and Ferrara (2016) laid out a synthesis of studies that involved comparative judgments of student performance and concluded that the assessment process could become more efficient using *adaptive* comparative judgment and the associated algorithm for determining pairs (L). This argument for ACJ over CJ, in terms of efficiency, resides in the pairs for comparative judgment being paired more optimally and efficiently—instead of having random comparisons, there are controlled, intentional pairings that makes the process more efficient in obtaining high levels of reliability (C. Rangel & M. Wingfield, personal communication, August 29, 2017). Bartholomew, Strimel, & Zhang studied the time required for teachers to make assessments through traditional approaches and through ACJ for open-ended assignments and found that the ACJ process took significantly more time than traditional approaches. However, they pointed out that these findings were not universal to all teachers in the study and the assessment tendencies of the teachers, in both ACJ and traditional assessment, were significantly different by teacher and grade. Overall, they concluded that between 8-10 rounds of judgment was required for a steady rank-order of student products to emerge in ACJ given the participants and artifacts in their research (S).

Pollitt and Whitehouse (2012) explored the rounds of judgement necessary for a reliable rank order. Their research showed that with the adaptive algorithm a reliability coefficient of 0.97 was reached after 12 rounds of judgment with additional rounds not contributing to a significantly higher reliability.

Validity and Reliability. Efforts towards determining the assessment validity, which refers to how closely the assessment actually measures the identified construct (Messick, 1992), of the ACJ output (e.g., the rank order of the artifacts included in the judgment session) have revolved largely around comparing the rank order with the results (e.g., scores) from traditional approaches to assessment (see Table 4).

Consistently, across the studies included, ACJ has produced high levels of reliability in the resulting rank order. The majority of studies included demonstrated levels of reliability $r \geq .8$ (see Table 4). The reliability coefficient produced in conjunction with rounds of ACJ assessment refers to the confidence of a subsequent session producing similar results given the session parameters. Many of the arguments in favor of using ACJ as an assessment tool for open-ended problems have largely centered on the high levels of reliability achieved through this approach (H, I, L, N, Q, R, &S).

Synopses of validity and reliability from the included research are included here:

H. Pollitt and Whitehouse (2012) looked at essays written from physical and human geography writing prompts. These results showed a correlation of 0.63 when compared to traditional marking with a reliability coefficient of 0.97.

I. Newhouse (2014) implemented ACJ in both Visual Arts and Design classes at a University level. They compared ACJ rankings with analytical marking and found a correlation coefficient of about 0.5. Additionally, they found reliability levels of 0.95 for Design and 0.93 for Visual arts.

L. Steedle & Ferrara (2016) looked at the validity and reliability of two writing prompts. When compared with traditional test scores, they found a correlation coefficient of 0.78 on Prompt 1 and 0.76 on Prompt 2.

N. Bartholomew, Reeve, Veon, Goodridge, Stewardson, Lee, & Nadelson (2017) used ACJ for the assessment both the portfolios and products of a middle school open-ended design challenge. A high interrater reliability was found for both the portfolios (r = 0.97) and the products (r = 0.96).

P. Bartholomew, Strimel, & Jackson (in press) compared the rank order of student portfolios, as obtained by a panel of judges with engineering and design background, with the scores obtained by student through traditional assessment approaches. Their analysis revealed a significant correlation.

Q. Seery, et al. (2017) implemented their study in an Initial Technology Teacher Education and showed high reliability on four separate assignments. The reliability on the four separate assignments was 0.97 each time.

R. Bartholomew, Strimel, & Yoshikawa (2017) compared the final rank order of middle school student design projects with the scores received by the students from their teacher using traditional assessment approaches. The results demonstrated a significant correlation suggesting validity around ACJ as an assessment approach.

S. Bartholomew, Strimel, and Zhang (2017) looked at the assessment of two open-ended design projects at an elementary level. Each teacher assessed the work of the students in their own class. All four teachers recognized the comparable results of rank order to traditional grading methods, however, because their study explored the use of a solitary judge no reliability was obtained/reported.

In addition to comparing the rank orders with traditional marking Bartholomew, Strimel & Jackson (in press) compared the ACJ rank order with the actual effectiveness of student prototypes in accomplishing the designed task. In this research the students were tasked with producing a water filtration device; the final turbidity levels of the provided water to each of the students were recorded as a representation of the effectiveness of their actual design. The analysis revealed that, while the rank order and the student scores received from their teacher were highly-correlated, the final turbidity level of the students' designs (e.g., the design effectiveness) and the ACJ rank order were not significantly correlated. This suggests potential issues with assessment validity as the actual effectiveness of the student designs were not represented in the final ACJ-produced output.

Table 4

Reliability and Validity Results across Studies

ID	Author	Validity (i.e., comparison with traditional grading)	Reliability level
C	Seery, Canty, & Phelan (2011)		*r* = .96
H	Pollitt & Whitehouse (2012)	*r* = 0.63 (with traditional marking)	α=0.97(interrater reliability)
I	Newhouse, C.P. (2014)	r ≈ .5 (with analytical marking)	*r* = .95 (design) *r* = .93 (visual arts)
L	Steedle & Ferrara (2016)	*r* = .78 (prompt 1) *r* = .76 (prompt 2)	*r* = . 89 (prompt 1) *r* = .78 (prompt 2)
M	van Daal, Lesterhuis, & Coertjens (2016)	SSR=.84	
N	Bartholomew, S.R., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., & Nadelson, L. (2017)		*r* = .96 (design products) *r* = .97 (design portfolios)
P	Bartholomew, Strimel, & Jackson (2017)	No significant correlation between ACJ rank or traditional rank and turbidity score. *r* = -.79 (ACJ rank and rubric score)	*r* = .95
Q	Seery, Delahunty, Canty & Buckley (2017)		*r* = . 0.974 (Assignment 1) *r* = . 0.973 (Assignment 2) *r* = . 0.965 (Assignment 3) *r* = . 0.971 (Assignment 4)
R	Bartholomew, Strimel, & Yoshikawa (2018)	*r* = -.56 (midpoint rank and rubric score) *r* = -.61 (final rank and rubric score)	*r* = .93 (midpoint) *r* = .97 (conclusion)
S	Bartholomew, Strimel, & Zhang (2018)	*r* = -.42	

The reliability level of the ACJ output is contingent on the rounds of judgment completed (G, I, & L). Pollitt (2012) suggested 12 rounds as a target for achieving a reliable rank order while Bartholomew, Strimel, & Zhang's (2017) suggested that a reliable rank order may be achieved as early as 6 rounds if ACJ were utilized by a solitary assessor (P). Steedle and Ferrara (2016) reported that reliability was above 0.8 by 9 rounds of judgment and slowly increased with further judgments (L).

Conclusion

Our analysis focused specifically on findings from studies on the recent uses and practices of ACJ in K-16 settings. Based on our synthesis we believe that ACJ demonstrates great potential for improving, informing, and potentially revolutionizing open-ended problem assessment in K-16 settings. With markedly higher reliability than other forms of assessment, coupled with other benefits in formative feedback, peer-review, and collaboration, ACJ is a potent tool in it's infancy of K-16 implementation. However, issues related to increased time investment requirements and assessment validity must also be noted and research into these topics will clarify the future potential for ACJ in educational settings.

This review outlined the research and findings around ACJ in K-16 education with the intention of informing decision-makers regarding the benefits, challenges, and research-based conclusions related to ACJ. Further, this review serves to highlight necessary future research areas related to settings, content areas, populations, and contexts within which the implications of ACJ-integration have not yet been explored. As the majority of ACJ research has taken place in college settings and in the areas of design, technology, and engineering education, we maintain that efforts towards broadening the context and samples included in ACJ-related research would further strengthen the understanding around the potential and implications for ACJ integration into K-16 education. Furthermore, the majority of ACJ research has emphasized summative assessment techniques and has shown positive results. Preliminary efforts into utilizing ACJ as a formative tool for assessment and feedback has shown promise and further efforts in both summative and formative applications of ACJ will strengthen arguments related to ACJ's potential for educational transformation.

From our review it was clear that the majority of research into ACJ for K-16 assessment has been conducted by a select group of researchers in a small sample of locations—most notably in the United Kingdom. We contend that future work into integrating ACJ in new settings, content areas, and with diverse samples will shed additional light and prove valuable in the conversation moving forward around ACJ for K-16 education.

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. (2017a). Assessing Open-Ended Design Problems. *Technology and Engineering Teacher*, *76*(6), 13-17.

Bartholomew, S. R. (2017b, November). Adaptive Comparative Judgment: Open-ended Problem Assessment and Student Learning. In *Mississippi Valley Technology Teacher Education Conference 104, St. Louis, Missouri.*

Bartholomew, S., Hartell, E., & Strimel, G.J. (2017). ACJ: A Tool for International Assessment & Collaboration. In *PATT34 Millersville University, Pennsylvania, USA 10–14 July, 2017*.

Bartholomew, S.R., Reeve, E., Veon, R., Goodridge, W., Stewardson, G., Lee, V., Nadelson, L. (2017). Relationships between access to mobile devices, student self-directed learning, and achievement. *Journal of Technology Education, 29* (1), 2-24

Bartholomew, S.R., Strimel, G.J., & Jackson, A., (2017). A Comparison of Traditional and Adaptive Comparative Judgment Assessment Techniques for Freshman Engineering Design Projects. *International Journal of Engineering Education, 34* (1), 20-33

Bartholomew, S.R., Strimel, G.J., & Yoshikawa, E. (2017). Using Adaptive Comparative Judgment for Student Formative Feedback and Learning During a Middle School Open-ended Design Challenge. *International Journal of Technology & Design Education.* https://doi.org/10.1007/s10798-018-9442-7

Bartholomew, S.R., Strimel, G.J., & Zhang, L. (2018). *Examining the Potential of Adaptive Comparative Judgment for Elementary STEM Design Assessment*. Manuscript submitted for publication.

Black, B., & Bramley, T. (2008). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23*(3), 357-373.

Borrego, M., Foster, M.J., & Froyd, J.E. (2014). Systematic literature reviews in engineering education and other developing interdisciplinary fields. *Journal of Engineering Education, 103*(1), 45-76. doi: 10.1002/jee.20038

Bramley, T. (2015). Investigating the reliability of Adaptive Comparative Judgment. *Cambridge Assessment, Cambridge, 36*.

Dearing, B. M., & Daugherty, M. K. (2016). Delivering engineering content in technology education: Can the technology education profession deliver on the promise of technological literacy for all while preparing the secondary school student for engineering education? *The Technology Teacher*, *64*(3), 8-12.

DigitalAssess (2015). *Comparative Judgment Update.* Retrieved on February 12, 2018 from http://digitalassess.com/comparative-judgement-update/

DigitalAssess (2017). *What we do: CompareAssess.* Retrieved on September 27, 2017 from http://digitalassess.com/what-we-do/#compareassess

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.

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).

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

Jones, I., & Alcock, L. (2012). Summative peer assessment of undergraduate calculus using adaptive comparative judgement. In P. Iannone & A. Simpson (Eds.), *Mapping University Mathematics Assessment Practices*. Norwich: University of East Anglia.

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

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

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

Kimbell, R., Wheeler, T., Miller, S., Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. London, England: TERU, Goldsmiths, University of London.

Kumar, M., & Natarajan, U. (2007). Alternative Assessment in Problem-Based Learning: Strengths, Shortcomings and Sustainability. *i-Manager's Journal on Educational Psychology*, *1*(1), 27.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *ETS Research Report Series, 1*(42), i-42.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International journal of qualitative methods*, *1*(2), 13-22.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229-258.

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.

National Research Council (NRC). (2009). *Engineering in K–12 education: Understanding the status and improving the prospects*. 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 (141-180). Sense Publishers.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: principles, policy & practice*, *21*(2), 205-220.

Pollitt, A. (2004, June). Let’s stop marking exams. In *IAEA Conference, Philadelphia*.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice, 19*(3), 281-300.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Studies in language testing, 3*, 74-91.

Pollitt, A., & Whitehouse, C. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

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

Sanders, M. (2009). STEM, STEM education, STEM mania. *Technology Teacher, 68(4)*, 20–26.

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.

Seery, N., Delahunty, T., Canty, D., & Buckley, J. (2017). Illustrating Educational Development Through Ipsative Performance in Design Based Education. In *PATT Conference, Philadelphia*.

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education*, *29*(3), 211-223.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286 (Reprinted as Chapter 3 from Thurstone, L. L. (1959). *The measurement of values*. Chicago, IL: University of Chicago Press).

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 1-16.

Wicklein, R. C. (2006). Five good reasons for engineering as the focus for technology education. *The Technology Teacher, 65*(7), 25.