International Journal of Technology and Design Education
https://doi.org/10.1007/s10798-020-09642-6

A systematized review of research with adaptive comparative judgment (ACJ) in higher education

Scott R. Bartholomew[1] · **Matthew D. Jones**[2]

Accepted: 23 November 2020

Abstract
Adaptive Comparative Judgment (ACJ), an approach to the assessment of open-ended problems which utilizes a series of comparisons to produce a standardized score, rank order, and a variety of other statistical measures, has demonstrated high levels of reliability and validity and the potential for application in a wide variety of areas. Further, research into using ACJ, both as a formative and summative assessment tool, has been conducted in multiple contexts across higher education. This systematized review of ACJ research outlines our approach to identifying, classifying, and organizing findings from research with ACJ in higher education settings as well as overarching themes and questions that remain. The intent of this work is to provide readers with an understanding of the current state of the field and several areas of potential further inquiry related to ACJ in higher education.

Keywords Adaptive comparative judgment · Design & technology · Assessment · Comparative judgment

Introduction

An emphasis on elevating current educational approaches to better prepare students for the future has often centered on open-ended learning scenarios (Viseu and Oliveira 2017; Munroe 2015; Hannafin et al. 1994). Learning approaches such as problem-based (Boud and Feletti 2013), project-based (Bell 2010), design-based (Bartholomew 2017), and inquiry-based learning (Lazonder and Harmsen 2016), have all been lauded, implemented, criticized, and studied towards that end (Duran and Dökme 2016; Kirschner et al. 2006; Thomas 2000; Mills and Treagust 2003). In tandem with these efforts towards open-ended learning, the struggles with accurately assessing them continues; the open-ended nature of these approaches to pedagogy often results in difficulties related to reliable, valid, and

✉ Scott R. Bartholomew
scottbartholomew@byu.edu

Matthew D. Jones
jone1947@purdue.edu

[1] Brigham Young University, 230 Snell, BYU, Provo, UT 84602, USA

[2] Purdue University, Young Hall, West Lafayette, IN 47907, USA

Published online: 01 January 2021

feasible assessment due to the wide variety of potentially "correct" answers (Bartholomew 2017; Kumar and Natarajan 2007).

Approaches to assessing these ill-structured, open-ended problems, have ranged widely with efforts around utilizing rubrics (Stevens and Levi 2013), presentations (Akister et al. 2000; Dobson 2006), and other criterion-based approaches taking center stage (Baartman et al. 2007). However, despite a large amount of research into these approaches, issues with reliability remain (Purzer et al. 2016; Bartholomew 2017; Moskal et al. 2002; Kimbell 2007). One alternative approach to these methods—specifically rising out of the need to improve reliability in open-ended assessment—is comparative judgment (CJ). Originally posited in 1927 by psychologist Louis Thurstone (1927), CJ centers on the idea that humans are intrinsically more skilled (e.g., reliable) at making judgments *between* items, through comparison, than they are at making value-laden judgments about the quality of an item (Thurstone 1927). In a CJ setting a judge views a pair of items and, rather than assigning values or scores to each item based on a scale, criteria, and rubric, simply chooses the item they believe is better. This decision is guided by a holistic criterion for judgment around which the decision should center (Pollitt 2012) and is intended to represent their understanding of a variety of facets related to the items displayed (Bartholomew et al. 2018b; Pollitt 2012). By repeating this process with successive pairings of items a markedly higher reliability level than traditional approaches to open-ended assessment can be obtained (Pollitt 2004, 2012) with the result of this CJ process being a rank order of all items included from the best to the worst.

Research into CJ assessment in educational settings has been documented elsewhere (Bartholomew and Yoshikawa 2018; Steedle and Ferrara 2016; Newhouse 2011; Verhavert et al. 2019) with many efforts in recent years emphasizing the ability of technology tools to facilitate and automate this process (Kimbell 2012a, b). Leaders in this field such as Pollitt and Murray (1996), Pollitt (2004, 2012), Kimbell (2012a, b), and Bramley (2015) have all added to our understanding of the potential—including strengths, weaknesses, reliability expectations, and other constraints—for this approach to improve assessment reliability. One specific effort to both improve the reliability and the efficiency of CJ has been the inclusion of an algorithm to *adaptively* select items for pairing based on previous pairing outcomes (Pollitt 2004). Both CJ and adaptive comparative judgment (ACJ) rely on Swiss-tournament rounds of judging but, in ACJ, a judge views pairs of items and both their choices, and the choices of other included judges, inform an algorithm which guides subsequent pairings; this intentional pairing through *adaptively-selected pairs* serves to both refine the resulting statistical outcomes and expedite the process of achieving a reliable output (i.e., standardized scores, rank order, and misfit statistics; Pollitt 2012; Rangel-Smith and Lynch 2018).

ACJ has been implemented in a variety of classroom settings, including as a tool for summative assessment, formative assessment, and student learning, with the majority of research being conducted in higher education settings. Further, implementation has spanned multiple locations across Europe, North America, and Australia (Bartholomew and Yoshikawa 2018). Additionally, the algorithm embedded in ACJ, which works to improve reliability, has been challenged (Bramley 2015), supported (Pollitt 2015) and refined (Rangel-Smith and Lynch 2018) and the body of ACJ-related research has continued to grow (Bartholomew and Yoshikawa 2018).

In 2018, Bartholomew and Yoshikawa synthesized the ACJ literature in K-16 educational settings. This review of the research covered a wide breadth of student ages, with limited synthesis and study depth reporting. The years following their review have shown the amount of research of ACJ conducted at the university level and beyond is greater in volume than in any

other educational setting, with this area continuing to grow at a rapid pace. Taken together, there is a need for review of ACJ-related research in higher education with an intentional focus on in-depth reporting of related studies. This review will serve those interested in understanding the research base around this field—specifically at the higher education level—as a starting block for future endeavors with an eye towards learning from past efforts. This synthesis will serve as a companion to other ACJ literature (e.g., Bartholomew and Yoshikawa 2018; Bramley 2015; Pollitt 2012; Verhavert et al. 2019), with a narrower focus (higher education ACJ implementation) and deeper analysis of processes, approaches, and findings from ACJ implementation in higher education. Additionally, in line with guidelines for systematized reviews (Grant and Booth 2009), commentary around potential strengths, weaknesses, and improvements for the reported work will be provided to assist in future research planning. The guiding question for this review was:

> **RQ**: What are the key findings related to research around the implementation of Adaptive Comparative Judgment in higher education settings?

Method

Systematized literature reviews

This effort was guided by the identified research question, previously conducted reviews (e.g., Bartholomew and Yoshikawa 2018), and guidelines provided by Borrego et al. (2014) and Grant and Booth (2009) for reviews of literature. Using these guides, we investigated ACJ-related literature specific to higher education settings. Our approach involved collecting research conducted on the topic, refining and narrowing the results, highlighting key findings guided by our research question, and providing a limited appraisal of the works reported including the findings, areas needing additional clarity, and potential bias or flaws in the design. Our intent is not to present every example of work related to ACJ; rather, we intend this review to provide a foundational "starting block" for those interested in ACJ in higher education (grade 13+, students ages 18+).

We include, at the beginning of our review, an informational section—as a point of reference and context—with findings from articles centered on the method of ACJ. These articles were identified by the authors during the review process and deemed as critical for providing insight and an understanding into ACJ and specifically the reliability of the approach—the most debated aspect of ACJ (Bramley 2015; Pollitt 2015; Rangel-Smith and Lynch 2018). These articles—which are included to provide context for the remaining articles included in the systematized review—include pieces which we determined to be "seminal" (having more than 75 citations at the time of this effort). The "seminal" articles were not all exclusively couched in higher education settings but their impact on future efforts and ACJ were deemed significant enough to warrant inclusion in the review to provide appropriate context to the reviewer.

Understanding adaptive comparative judgment

Using the stated criteria, several articles were identified and deemed critical in situating ACJ as an assessment approach; these provide context and understanding to the approach and often centered on the trustworthiness of the reliability measures reported. While Pollitt (2012) utilized research with young student writing samples to argue the reliability of the ACJ method, other articles (Pollitt 2004; Thurstone 1927) were mainly theoretical in nature with an emphasis on statistical and procedural methods employed. Bramley (2015) provided evidence to suggest the reliability measures reported through ACJ as inflated and invalid. However, Pollitt's (2015) criticism of the nature of Bramley's (2015) methodology responded and worked to resolve these concerns. More recent efforts by Rangel-Smith and Lynch (2018) have further responded to the chief criticisms of Bramley's (2015) and Pollitt's (2015) work, offering additional modifications to the algorithm and approach that would raise trust in the reliability measures reported. The majority of these stage-setting articles did not employ research with ACJ in educational settings with students—rather, modeling through statistics and computer functions were employed. Two additional articles summarized below (Bramley and Vitello 2018; Benton and Gallagher 2018) also focused on the reliability of the ACJ approach; these articles included research in educational settings—as opposed to being limited to statistical modelling—and, as such, will be included in the later analysis. Specific to our purpose in this review, we did not include articles focused on CJ, as opposed to ACJ, or articles centered on the development of the software platform or training of assessors. Researchers interested in ACJ should review the provided summaries of these pieces, in addition to the other items reported, as we perceive them as important to the understanding of ACJ and its place with educational settings. To set the stage for our review we provide a brief synopsis of these articles and the associated findings here:

1. Thurstone (1927) lays the foundation for "a new psychophysical law which may be called the law of comparative judgment (p. 273)." Thurstone argues that CJs between items can be used to identify which item is "stronger (better, lighter, more excellent) than [another] (p. 285)" and thus reliably create a rank ordering of a group of items from highest quality to lowest. Thurstone demonstrates the applicability of this law for one judge, or a group of judges, and highlights the potential of this approach to judging "handwriting specimens, children's drawings, or any other series of stimuli that are subject to comparison (p. 273). Thurstone's article, which has been cited more than 6500 times, provided the foundation for future comparative judgment work.
2. Pollitt (2004) argues for CJ—as opposed to traditionally-used marking (assessing) methods—which he contends are flawed with bias, reliability, and validity issues. In this article, which has been cited almost 100 times at the time of this review, Pollitt points out that traditional approaches utilize a

> summative assessment system in which judgment of the quality of the achievement of students is compromised by an unnecessary concern for the reliability of marking by a widely distributed team of markers whom it is difficult and expensive to monitor well. Yet the alternative of paired comparative judgment has been shown to work adequately with these same current exams… a paired comparison (or rank ordering) system would provide much more precise quality control over the essential processes of evaluating students' performances (p. 20).

Pollitt specifically points out that advances in computing make possible what once was not (i.e., scanning, uploading, and web-based dissemination of student work for judging) and that the time was right for adopting Thurstone's suggestions and using CJ in summative assessment for subjective work.

3. Pollitt (2012) contains the findings from several trial studies with ACJ—a modification of Thurstone's (1927) proposed CJ approach. In this article, with over 150 citations at the time of this review, Pollitt argues that professional judgment by teachers—applied through ACJ in series of comparisons—is more reliable than traditional approaches and yields valid results. Pollitt outlines the process of *adaptively* pairing items to refine the final ranking and parameter values and details "quality control" measures inherent in the process. Pollitt also notes lurking questions around how judges make judgments in different contexts and readily notes that such an approach to assessment may not be accepted by the public. Pollitt ends by arguing (p. 20):

> We currently operate a summative assessment system in which judgment of the quality of the achievement of students is compromised by an unnecessary concern for the reliability of marking by a widely distributed team of markers whom it is difficult and expensive to monitor well. Yet the alternative of paired comparative judgment has been shown to work adequately with these same current exams. Future exams could be designed to meet the summative purpose directly—and changes of this kind would only make things better for a judgment system.

4. Bramley (2015) used two computer simulations, both using the Rasch formulation model for comparisons, to measure the validity of the reliability measures of ACJ. The researcher concluded that the "SSR statistic is…misleading…as an indicator of scale reliability whenever a CJ study has involved a significant amount of adaptivity" (p. 15). Bramley discusses this "inflated" reliability produced through ACJ and contends against trusting the results of the approach in terms of reliability levels.
5. Pollitt (2015), responded to Bramley's (2015) article with the report of his own Rasch modeling statistic simulation of ACJ. Pollitt argued that ACJ measures internal consistency, not necessarily reliability—something he terms a "judge consistency coefficient" (JCC). Pollitt notes some "upward bias" in the scoring of consistency in ACJ, but also contends that this is "extremely small…practical[ly] negligible" (p. 1).
6. Rangel-Smith and Lynch (2018) responded to Bramley's (2015) and Pollitt's (2015) articles by analyzing the algorithm underpinning the adaptivity in ACJ. They developed coin-flip/mathematical simulations and compared these to ACJ in an effort to specifically investigate Bramley's (2015) challenge of the reliability reported through ACJ. Their results showed some bias in the adaptivity of ACJ which confirmed Bramley's arguments. However, the authors reported several steps which were undertaken to mitigate and/or eliminate these issues in the algorithm moving forward.
7. Bramley and Vitello (2018) performed two studies: the first looked at ACJ compared to an All-Play-All method. The second looked at Random Comparative Judgment (RCJ). Their results led them to conclude that "the SSR statistic should not be used as the basis for drawing strong conclusions about reliability when adaptivity has been used" (p. 53). Their findings, which are similar to Bramley's (2015) arguments center on a lack of trust in the reliability values produced through ACJ. However, the authors also contend that future research on adaptivity and reliability needs to continue to fully understand the outcomes and implications of the ACJ findings and reliability levels.

8. Benton and Gallagher (2018) used a meta-analysis approach in combining existing data and performing their own pseudo-CJ simulation compared to traditional grading to render a correlational analysis. The purpose of the study was to discover if comparative judgment (particularly ACJ) was simply a form of multiple (repeated) marking. The authors contended against ACJ as an approach to expediting grading or improving reliability; specifically they noted that "the evidence from this study supports [our study's purpose]…. The physical act of placing two essays next to each other and deciding which is better does not appear to produce judgments that, in themselves, have any more predictive valuethan getting the same individual to simply mark a set of essays" (p. 6).

A review of these articles, which provide the backdrop for a clearer understanding of ACJ, made it clear that the debate over ACJ—including the approach, method, and statistical procedures—is ongoing. Reliability measures reported through ACJ are especially contentious with arguments for and against. The reported contradictory findings suggest the need for further inquiry into the feasibility of ACJ as an approach to both assessment and learning. Despite the debate over ACJ, and the reliability levels especially, there are several notable instances of significant positive impacts from ACJ implementation. Increasing the reliability of assessment approaches, creating an "easier" task for assessors, and the potential for learning to occur through ACJ all point to a potentially-positively-impactful opportunity; an opportunity that needs further investigation to more fully understand.

Systematized review: search parameters

Building on other synthesis work (Bartholomew and Yoshikawa 2018), we began the systematized review process by identifying and narrowing potential articles for inclusion. However, unlike Bartholomew and Yoshikawa (2018) our focus for collecting research was both narrower and broader in scope; narrower in our focus of ACJ research in university or higher educational settings (instead of K-16) and broader in our discrimination of what defined an educational setting (i.e. medical school, graduate studies, etc.). We initially identified ACJ articles through a keyword search ("ACJ," "Adaptive Comparative Judgment") which centered on ACJ as an approach; these articles, which we have included in the previous section, center on the underpinning method and mechanism of ACJ (e.g., discussions around reliability, validity, the adaptivity algorithm, etc.) as opposed to the findings of a research study which utilized ACJ. Our next step was to combine these stage-setting articles with an additional search for "seminal" articles; with the main criteria for inclusion as a seminal piece being more than 75 citations (according to Google Scholar, December 2019). These publications were summarized in the preceding section and, along with citing works, were used to further refine our search parameters.

After investigating and summarizing the articles which set the stage for ACJ, we proceeded to expand our search; this was conducted using the following key words: (1) Adaptive Comparative Judgment, (2) ACJ, (3) Comparative Judgment, and (4) CJ. Boolean phrases such as "Adaptive Comparative Judgment AND/OR ACJ", "Comparative Judgment AND/OR CJ" were also used. These terms were selected given their connection to, and use with, ACJ related research. The authors noted that these terms were sometimes used interchangeably—even in instances describing something that did not fit both terms. Given the potential overlap in search terms, and the confusion of their use in the past, the authors deemed it necessary to use these four terms independently in

searches to glean all potentially relevant articles. Each keyword search was conducted in several academic search engines including Google Scholar, ERIC via EBSCOhost, and Education Full Text.

These four databases were utilized in an effort to "cast a wide net" and identify as many potential articles as possible for inclusion. Google Scholar is not an isolated database only showcasing flagship journal articles but is far more open and networked than other research databases and thus may yield innovative research methodologies related to meta-analyses and other literature reviews (Zientek et al. 2018; Martin-Martin et al. 2017). Using such a database provided us an advantage in conducting this review of the extant literature because it allowed us to "trace interconnections among authors citing articles on the same topic" and to determine the frequency of citation (Noruzi 2005, p. 170; see also Zientek et al. 2018). As reporting the number of citations was key to our understanding of what constituted a "seminal" article (Martin-Martin et al. 2017) this database was used. ERIC is a database featuring academic research in educational settings and as our purpose was to analyze ACJ in educational settings this research tool was considered and used. Education Full Text was recommended by the researching institution's librarian as another database to be used in the literature review because of its expansive inclusion of educational research and added features of including easy access to copies of the full text articles under consideration.

The aforementioned search terms entered into the databases yielded, in total, over five million results on Google Scholar, 2371 results on ERIC via EBSCOhost, and 146 results on Education Full Text. A review revealed that the majority of these articles were not related to our research question; therefore, further refinement of searching included searching for "Adaptive Comparative Judgment" with the addition of parentheticals, and removal of duplicate articles. This refinement resulted in a total of 55 articles selected for further analysis. In addition to this online searching effort, contact was made with leading individuals in ACJ soliciting other related research and items of potential relevance. While this effort resulted in many additional articles, only 47 were included in the next step of refinement based on meeting the identified criteria. This resulted in a total of 102 items (55 + 47).

After identifying 102 items, our next step in refining the potential articles was the removal of duplicates—this included both replicated works and items which were presented in multiple formats (i.e., conference presentation/paper and peer-reviewed publication). Following the removal of duplicates, the collected works were classified and either removed or included based on several predetermined criteria. This process of classification and removal/inclusion followed a review of abstracts, introduction, methods, and findings sections for each article. The classification categories, and subsequent criteria for inclusion, are included here:

1. **Higher education context**: the context of the paper should be limited to studies around ACJ in education settings at a higher education level. Although termed differently in different locations, "higher education" is interpreted to mean post-secondary schooling occurring at a university or college. These settings include grades following 12th grade with students 18 years old and older.
   (a) **Not included**: papers with research involving settings outside of higher education such as K-12 education, workforce development, or business settings.
2. **Original research**: all included works should reflect original findings and research around the use of ACJ.

(a) **Not included**: papers from the same author without new findings or explanations such as conference papers which were later published in journal articles or multiple articles using the same data.
(b) **Not included**: "works in progress" or papers which had not been peer-reviewed (i.e., white papers).
(c) **Not included**: publications about "Project e-scape." Although many of these articles could be considered "seminal" in terms of citation count, the setting of this project was not higher education and many of these articles revolve around the creation of software and design tools rather than findings from ACJ implementation.

3. **Application**: articles must focus on the application of ACJ in an educational setting.
   (a) **Not included**: articles centered on a discussion around the method of ACJ (i.e., how items are paired, the reliability of the algorithm, and so forth).

Following the review, classification, and winnowing process, 21 articles met the identified criteria and were selected for this systematized review. The majority of articles not selected, from the 102 identified, were removed for not meeting the following criteria: must be studying ACJ (not simply CJ); must be in higher educational settings; must represent peer-reviewed work; and must present original empirical findings from research. An overview of our selection approach can be seen in Fig. 1. Table 1 includes the articles selected for inclusion in our meta-synthesis.

ACJ for research: mapping the results

Turning attention to the articles focused on educational research with ACJ, an initial scan shows a clear trend towards increasing adoption of, or at least investigation into, the use of ACJ in higher education settings. With the exception of 2012—during which one edition of the *International Journal of Technology & Design Education* had a special issue (vol. 22) around ACJ—there has been a continuous rise in research articles and published conference proceedings related to ACJ (see Fig. 2). It should be noted that at the time of this publication preparation (Fall 2019) the year 2019 was not completed.

Context

While the theoretical underpinnings of ACJ began in the United States of America with noted psychologist Louis L. Thurstone (1927), the initial efforts into implementing ACJ in educational settings were largely confined to researchers in the United Kingdom (Pollitt, Jones, Alcock, Kimbell, and Bramley) and Ireland (Seery, Lane, Canty, Buckley, Doyle, Phelan, and Rowsome). All the research articles published between 2004 and 2015 came from these two geographic locations until 2016 when Metzgar, from the United States, published his work related to ACJ for an MBA course. From 2016 to 2019 the geographic locations represented in the ACJ research spread to include additional research in the United States of America as well as locations such as Canada (Potter et al. 2017). However, the three epicenters of ACJ-related work in higher education appear to currently be the UK, Ireland, and the United States of America (see Fig. 3).

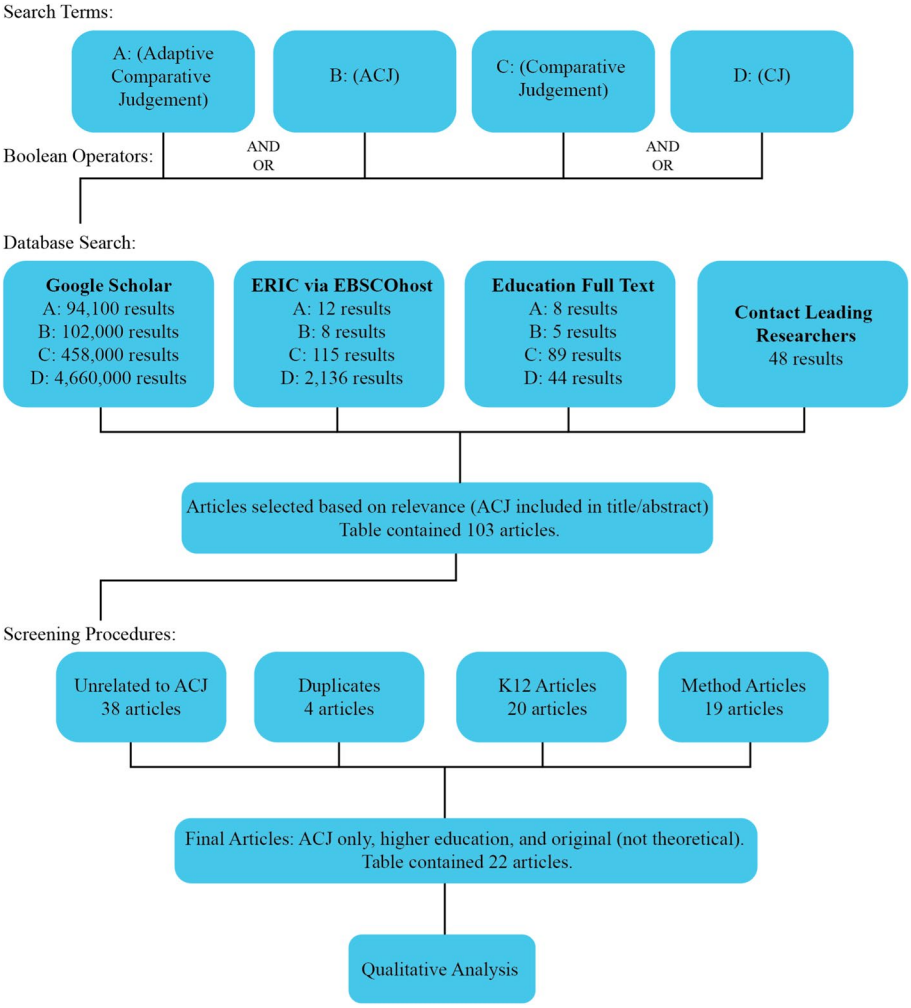

Fig. 1 Overview of systematized literature review

The participants included in ACJ-related research have almost exclusively been university students (see Fig. 4). Other assessors included in ACJ research have included industry professionals (4), university instructors/professors (6), and/or other academic staff (e.g., course TAs). The majority of research has situated university students as the assessors in ACJ research (21). In most of these instances the students have acted as assessors of their own work, their peers work, or similar work to that completed in the course (see Table 2).

In terms of subject area, the ACJ research in higher education has spanned a variety of courses including writing, mathematics, and design and technology (in Europe) or Technology & Engineering Education (USA). The most common subject matter for ACJ-assessment has been design portfolios situated in design and technology/Technology & Engineering Education courses (see Fig. 5). Not surprisingly, this finding mirrors the content area of the researchers most heavily involved with ACJ research (see Table 1).

Table 1 Articles selected for inclusion in the synthesis of ACJ-related literature at the higher education level

ID	Title	Author(s)	Source	Year
A	Exploring the value of democratic assessment in design based activities of graphical education	Seery, N., Lane, D., & Canty, D.	*118th Annual American Society of Engineering Education Conference*. Vancouver, British Columbia: American Society for Engineering Education	2011
B	Summative peer assessment of undergraduate calculus using adaptive comparative judgment	Jones, I. & Alcock, L.	In P. Iannone & A. Simpson (Eds.), *Mapping University Mathematics Assessment Practices*. Norwich: University of East Anglia	2012
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	2012
D	The impact of holistic assessment using adaptive comparative judgment of student learning	Canty, D.	PhD Thesis, University of Limerick, Ireland	2012
E	The development of pre-service design educator's capacity to make professional judgments on design capability using adaptive comparative judgment	Rowsome, P., Seery, N., & Lane, D.	American Society for Engineering Education	2013
F	The validity and reliability of Adaptive Comparative Judgments in the assessment of graphical capability	Seery, N., Buckley, J., Doyle, A., & Canty, D.	In *Proceedings of the 71st Mid-Year Conference of the Engineering Design Graphics Division* (pp. 104–109)	2016
G	Using adaptive comparative judgment to assess student work in an MBA course	Metzgar, M.	*International Journal for Infonomics*, 9(3), 1217–1219	2016
H	Integrating Peer Assessment in Technology Education through Adaptive Comparative Judgment	Canty, D., Seery, N., Hartell, E., & Doyle, A.	*In PATT34 Technology & Engineering Education–Fostering the Creativity of Youth Around the Globe, Millersville University, Pennsylvania, USA*, 10–14	2017
I	ComPAIR: A new online tool using adaptive comparative judgment to support learning with peer feedback	Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I.	Teaching & Learning Inquiry, 5(2), 89–113	2017
J	Adaptive Comparative Judgment: A Tool to Support Students' Assessment Literacy	Rhind, S. M., Hughes, K. J., Yool, D., Shaw, D., Kerr, W., & Reed, N.	*Journal of veterinary medical education*, 44(4), 686–691	2017

Table 1 (continued)

ID	Title	Author(s)	Source	Year
K	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	2018
L	Five Go Marking an Exam Question: The Use of Adaptive Comparative Judgment to Manage Subjective Bias	Barber, J.	*Practitioner Research in Higher Education*, 11(1), 94–100	2018
M	Exploring the Potential for Identifying Student Competencies in Design Education through Adaptive Comparative Judgment	Bartholomew, S.R., Yoshikawa, E., & Connolly, P.E.	In *PATT35 Athlone Institute of Technology, Athlone, Ireland 18–21 June, 2018,* pp. 187–194	2018
N	Integrating learners into the assessment process using adaptive comparative judgment with an ipsative approach to identifying competence-based gains relative to student ability levels	Seery, N., Buckley, J., Delahunty, T., & Canty, D.	*International Journal of Technology and Design Education*, 1–15	2018
O	A Tool for Formative Assessment and Learning in a Graphics Design Course: Adaptive Comparative Judgment	Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J.	*The Design Journal*, 1–23	2019
P	First, Last, Elsewhere… Positioning Adaptive Comparative Judgment in the Design Learning Experience	Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J.	*In PATT 37 University of Malta, Malta, June 2–6, 2019,* pp. 65–74	2019
Q	Student Peer Assessment Using Adaptive Comparative Judgment: Grading Accuracy versus Quality of Feedback	Demonacos, C., Ellis, S., & Barber, J.	*Practitioner Research in Higher Education,* 12(1), 50–59	2019
R	Inducting ITE students in assessment practices through the use of comparative judgment	Canty, D., Buckley, J., & Seery, N.	In *PATT 37 University of Malta, Malta, June 2–6, 2019,* pp. 117–124	2019
S	Investigating Differences in Formative Critiquing between Instructors and Students in Graphic Design	Zhang, L.	Doctoral dissertation, Purdue University Libraries	2019
T	Learning by Evaluating (LbE)	Bartholomew, S.R., Mentzer, N., & Jones, M.	In *Mississippi Valley Technology Education Conference*. Nashville, TN	2019

Table 1 (continued)

ID	Title	Author(s)	Source	Year
U	Informing Engineering Design through Adaptive Comparative Judgment	Strimel, G.J., Bartholomew, S.R., Purzer, S., Yoshikawa, E., & Zhang, L.	*European Journal of Engineering Education*. https://doi.org/10.1080/03043797.2020.1718614	2020[a]
V	An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education	Buckley, J., Canty, D., & Seery, N.	*Irish Educational Studies*. https://doi.org/10.1080/03323315.2020.1814838	2020[a]

[a]These articles were under review at the time of this publication preparation; during the subsequent revision process the references were updated to the 2020 publication information for each article, which was published online during the 2020 year

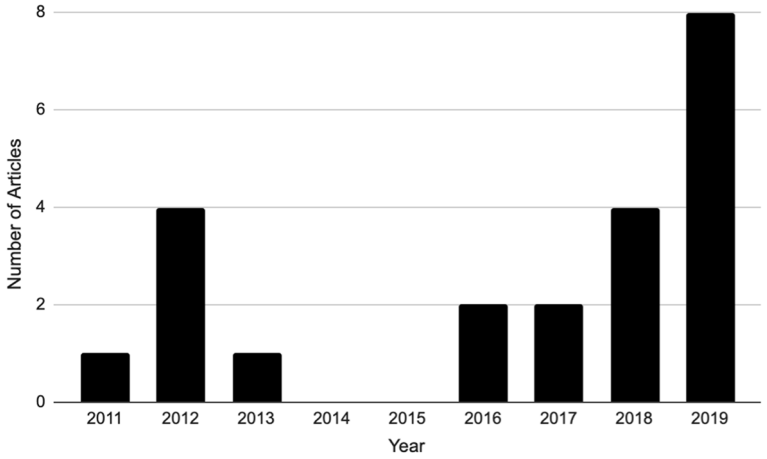

Fig. 2 The number of ACJ articles by year. *Note:* As the Strimel et al. (2020) and the Buckley et al. (2020) articles were in the peer-review process at the time of publication preparation they were included as 2019 articles (both submitted in 2019) in the figures in this systematized review. These articles have since been published online, during the 2020 calendar year, thus leading to the 2020 citation for each in Table 1

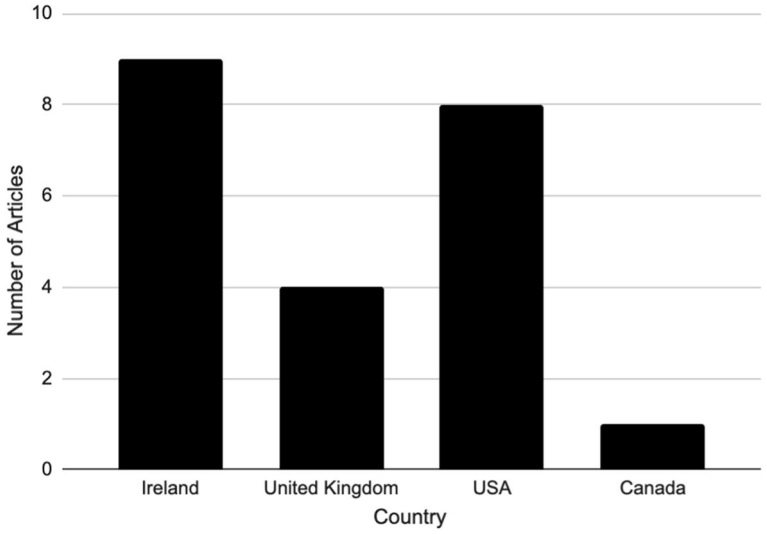

Fig. 3 Locations where ACJ research has been conducted (2016–2019)

As ACJ was founded on the idea that assessment in open-ended scenarios is difficult, unreliable, and time-consuming (Kimbell 2012a; Pollitt 2012), we expected to see student responses to open-ended assignments most commonly in our review of items included. This expectation was realized as the most commonly assessed items, through ACJ, were portfolios from design settings and projects (13), followed by writing assignments (6). Other items assessed through ACJ included math assignments, presentations, audio files, and other data formats (see Fig. 6).

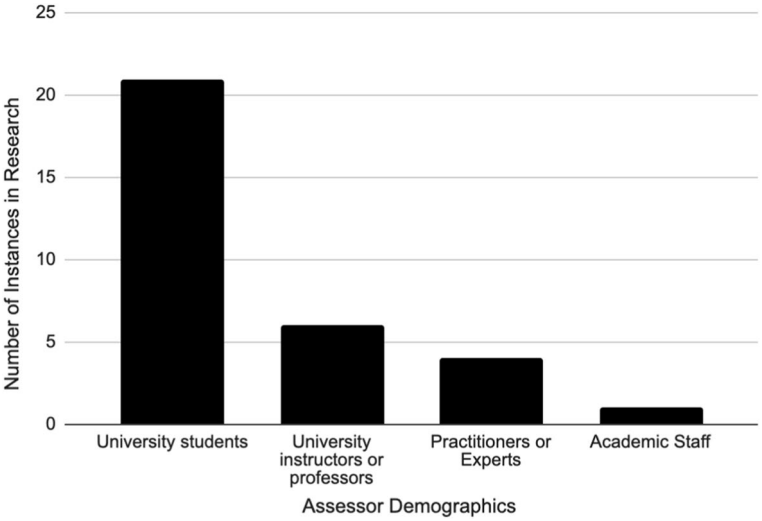

Fig. 4 Assessor demographics and the number of instances they appear in the research

ACJ implementation approaches

Thurstone (1927) helped set the foundation for studies in comparative and ACJ when he argued that CJs were more reliable than subjective judgments due to the difficulty in quantifying (e.g., assigning values) the quality of work. His work concerning a single observer with paired comparisons would not be made adaptive and inclusive of multiple observers until the early part of the twenty-first century (Pollitt 2004) when Pollitt presented the method of using CJ for summative assessment to increase reliability through multiple raters. Later, Pollitt further described (2012) the method of adaptivity in CJ and established the means by which many other researchers in other academic disciplines could utilize CJ/ACJ in K12 and higher educational settings.

The ACJ research in higher education has revolved mainly around summative (11) assessment interventions, however other approaches, including studies using formative or both formative and summative approaches to the study of ACJ (10), have also been prevalent. The placement of ACJ in the student learning experience has been the subject of some debate (Bartholomew et al. 2019b) in recent years with some contending the real value in ACJ may reside in assisting students in learning rather than simply improving reliability (Bartholomew et al. 2019c). Relative to these academic discussions around summative or formative use of ACJ we have provided a brief synopsis of each of the relevant articles here for reference. Additionally, we have added commentary around the strengths, weaknesses, and relevant opportunities for additional or strengthened research.

ACJ for Summative Assessment These studies specifically used ACJ in a summative assessment approach. Though some studies below used surveys as part of their research design, their inclusion here indicates the opinion of the researchers that these reflected a main focus on researching ACJ as a summative assessment tool.

Table 2 Higher education ACJ integration: settings, assessors, subject area, and participants

ID	Author (year)	Participants [n] & artifacts	Assessor demographics	Research location	Subject area
A	Seery, N., Lane, D., & Canty, D. (2011)	University students [121] Design portfolios	University students	Ireland	Materials & Construction Education teacher education; Materials and Engineering Teacher Education
B	Jones, I. & Alcock, L. (2012)	University students [169] Math scripts	University students	United Kingdom	Calculus
C	Seery, N., Canty, D., & Phelan, P. (2012)	University students [137 participating with 63 as judges] Student portfolios	University teachers	Ireland	Materials & Construction Education teacher education; Materials and Engineering Teacher Education
D	Canty, D. (2012)	University students [137, 133, 136] Design portfolio	University student peers and final round of experts	Ireland	Materials & Construction Education teacher education; Materials and Engineering Teacher Education
E	Rowsome, P., Seery, N., & Lane, D. (2013)	University students [13] Design portfolios	University students	Ireland	Materials & Construction Education teacher education; Materials and Engineering Teacher Education
F	Seery, N., Buckley, J., Doyle, A., & Canty, D. (2016)	University students [128] Design portfolios	University students	Ireland	Initial Technology Teacher Education
G	Metzgar, M. (2016)	MBA students [34] Writing scripts	University students	USA	Business
H	Canty, D., Seery, N., Hartell, E., & Doyle, A. (2017)	University students [136] Design portfolios	University students	Ireland	Initial Technology Teacher Education
I	Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I. (2017)	University students [approx. 450] writing scripts (English); writing scripts (Physics); drawing/diagram (Math)	University students	Canada	English, Math, & Physics
J	Rhind, S. M., Hughes, K. J., Yool, D., Shaw, D., Kerr, W., & Reed, N. (2017)	University students [154] Writing scripts	University students	United Kingdom	Veterinary medicine

Table 2 (continued)

ID	Author (year)	Participants [n] & artifacts	Assessor demographics	Research location	Subject area
K	Bartholomew, S.R., Strimel, G.J., & Jackson, A. (2018)	University students [16] Engineering notebooks, prototypes, and performance results	5 independent judges (experience in evaluating engineering design projects)	USA	Engineering
L	Barber, J. (2018)	University students [61] Writing scripts/test question	12 academic staff for 1st study; university peers for 2nd study	United Kingdom	Pharmaceutical
M	Bartholomew, S.R., Yoshikawa, E., & Connolly, P.E. (2018)	University students [28, 16, N/A, 36, 111] Portfolios (visual and/or audio)	University students, professors, practitioners	USA	Transdisciplinary Studies in Technology, Theater, Interior Design, Computer Graphics Technology, Engineering Education
N	Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2018)	University students [128] Design scripts	University students	Ireland	Initial Technology Teacher Education
O	Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J. (2019b)	University students [85] Graphic design drafts	University students and instructors	USA	Graphic Design
P	Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J. (2019c)	University students [158] and instructors [6] Design scripts	University students	USA	Graphic Design
Q	Demonacos, C., Ellis, S., & Barber, J. (2019)	University students [130] Writing scripts	University students	United Kingdom	Pharmaceutical
R	Canty, D., Buckley, J., & Seery, N. (2019)	University students [59] Assessment design portfolios4	University students	Ireland	Initial Technology Teacher Education
S	Zhang, L. (2019)	University students [3] and instructors [3] Graphic design projects	University students and instructors	USA	Graphic Design
T	Bartholomew, S.R., Mentzer, N., & Jones, M. (2019)	University students [550], and instructors [6] Design portfolios	University students and instructors	USA	Design and Technology
U	Strimel, G.J., Bartholomew, S.R., Purzer, S., Yoshikawa, E., & Zhang, L. (2020)	University students [110] Design presentations	University students, instructors, and practicing engineers	USA	Engineering

Table 2 (continued)

ID	Author (year)	Participants [n] & artifacts	Assessor demographics	Research location	Subject area
V	Buckley, J., Canty, D., Seery, N.	University students [126]	University students	Ireland	Initial Technology Teacher Education

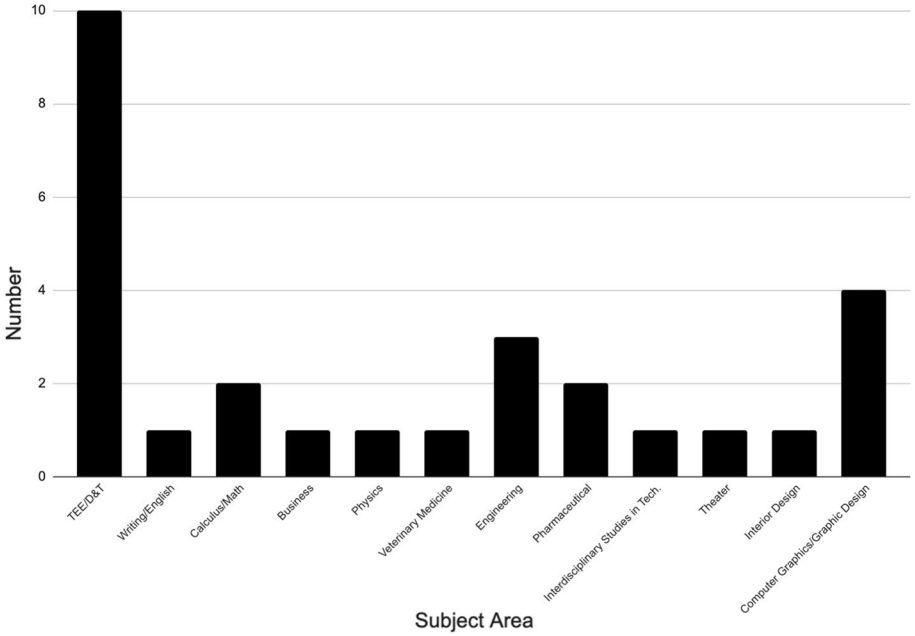

Fig. 5 Subject areas in which ACJ research has been conducted

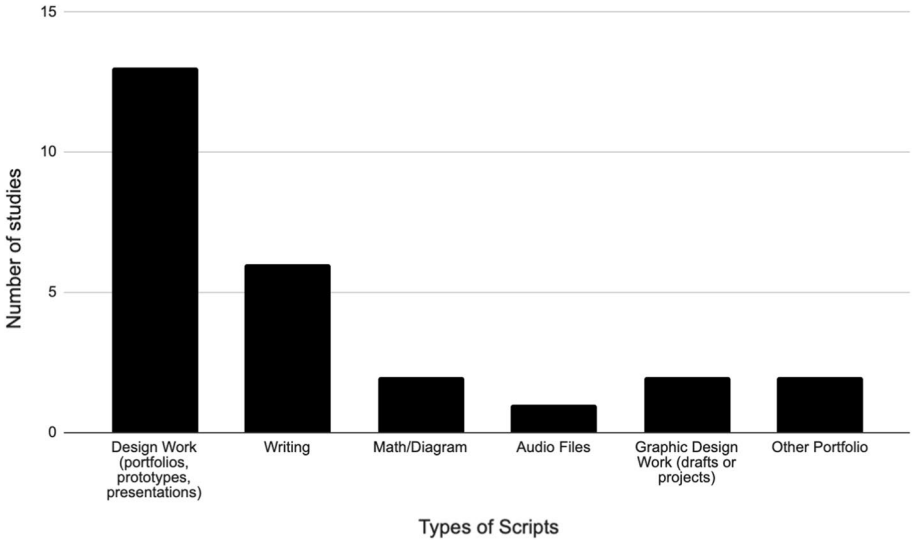

Fig. 6 Types of scripts used in ACJ research

A. Seery et al. (2011) had university students perform a design activity followed by making CJs on other student works. Their findings included that students have the capacity to identify evidence of learning and "to judge the value of analytical thinking even if they did not achieve this themselves" (p. 17). These findings suggest the

potential for ACJ as a learning and evaluation tool—even with "novice" students. However, a correlation between the rank order—produced by student judges—and the assessment results obtained from an instructor is not shared, thus indicating is a potential for the student judges (novices) to collectively obtain an "incorrect" (i.e., different from the instructor "expert") ranking of work.

B. Jones and Alcock (2012) used ACJ as a summative assessment tool with students, experts, and novices. They also conducted interviews and issued surveys. They compared inter-rater reliability among the three groups; they found that "experts and peers were more in agreement with one another about what constitutes a good answer to the question than were the experts and novices" (p. 68). However, the authors similarly noted that peers and novices were also significantly correlated in their responses—a counter-intuitive finding to the first. While the initial findings suggest the capacity of students to accurately perform judgments in ACJ settings—even in settings where they may not have achieved an "expert" understanding of the content—the additional finding of a high correlation between novices and peers suggests the need for further investigation. The authors point to this phenomenon as well as several others (e.g., judgment criteria were not provided to students suggesting the potential for achieving reliable rankings without an agreed upon criterion for judgments; because this research utilized a problem for which advanced mathematical skills are required an understanding of how the question and context impact the results is needed). We agree with the suggestions provided by the authors for additional research and link their findings to Strimel et al.'s work (2020) which similarly demonstrated differences in ACJ results between groups of assessors. Understanding these differences and the implications around them is needed for ACJ implementation to continue to spread.

C. Seery et al. (2012) had university students peer-grade student-made design projects. In this research ACJ was validated as being an effective summative assessment tool when compared to other assessment methods. Similar to Jones and Alcock (2012), the students were not assigned "predefined criteria" to make the judgments. The authors noted benefits to this approach such as the co-creation of the assessment criteria by both the assessors (students) and the assessee (students) the democratic process that resulted, and a closeness amongst students which was before unseen. Further, a significant correlation between the student-created order and that obtained from module leaders suggests validity in the results obtained by the students and strengthens arguments for the feasibility of this approach.

D. Canty (2012) used ACJ with students for assessment at the conclusion of a design task. Qualitative surveys were also taken of student performance, motivation, knowledge and skills, and other attributes. Students were able to assess "capability" having had no predefined criteria and ACJ was validated as an effective method for design-based education. Canty's work used a hybrid of traditional (e.g., rubric-based), ACJ, and qualitative data collection; the results showed high reliability and validity while using ACJ with student assessors. While the results from both the student-produced and the expert-produced ACJ sessions were significantly correlated, Canty noted (p. 157) that:

> the student and professional assessors agreed on the general location of a portfolio on the rank order but it is evident that within these regions of the rank order there is variance on the perceived quality of the work.

The implications of this variance within significantly correlated rank-orders is intriguing and further effort in this vein could produce promising insight into differences between student and expert perceptions.

E. Rowsome et al. (2013) had university students complete a design assignment and then use ACJ to assess the class projects. A non-invasive interview process was used to collect judges' thoughts as they completed the judgments. The results (e.g., reliability) were encouraging and the researchers concluded that student teachers could effectively assess student work as well as professionals, even when there were no predefined criteria before completing ACJ. They noted an added benefit in that judges became more confident in their decisions as they worked through consecutive judgments. Further, they noted that (p. 9):

> If both portfolios were impressive the judge would have to go down the list of criteria. There is reason to suggest that the participants had some hierarchical structure on their criteria.

This inherent "hierarchical structure" is worthy of further inquiry as the judges' decisions may have been influenced in specific ways or through different experiences. What determined the hierarchical structure and how this varied between students would shed additional light and perspective on the mental processes involved with ACJ.

F. Seery et al. (2016), gave students a design project that was followed by student-made judgments of work using ACJ. Students also hand graded (on a scale of 1–10) select portfolios for points. A correlational analysis was run between final points and ranks. The authors argue that high reliability levels demonstrate the potential and feasibility of ACJ; however, they also noted several correlations, between ACJ outputs and traditional marking scores, which were only moderate. The authors posit that the holistic nature of ACJ introduces the chance of other variables—not represented on the traditional marking rubric—being used to evaluate student work. What these "other" variables may be, how they vary amongst students, and how they may impact the ACJ experience and decisions is unknown. A fuller understanding of these issues may lend additional credence to the ACJ approach and how it impacts student experience.

G. Metzgar (2016) used ACJ with students at the end of each of five writing assignments. In each case the instructor also scored and rank-ordered all the assignments independently. At the end of each writing assessment the student produced ACJ rank order was compared to the instructor rank order and point values. In general, the student-produced and instructor-generated rankings were similar, however, as the author notes:

> one area of high agreement was the sorting of top performers and low performers. In four of the five tasks, the top three and bottom three were identical for the ACJ rankings and the instructor rankings. In all five tasks, the bottom three candidates were exactly the same for ACJ and the instructor (p. 2).

Meztgar also notes that, despite these similarities in high- and low-ranking items, there was a level of disagreement when it came to the mid-range rankings. This disagreement becomes more important as the author noted that the process of applying the ACJ-produced value to a grade was a "subjective process (p. 3)." Understanding the nuances of the differences in mid-range rankings may be both insightful and important—especially if the output of ACJ judgments will be used to determining student grades.

H. Canty et al. (2017) used ACJ as a summative assessment tool with university students. The researchers incorporated the use of surveys to gauge student experiences with the democratic process of "establishing what to value (p. 4)" through the ACJ

assessment process. The authors reported that the university students developed not only the ability to assess other student work with high reliability, but also the capacity to appraise their own standards and performance and engage in the subject domain for themselves. In this instance the students reacted positively to the removal of explicit criteria for assessment and "indicated that this empowered them to be innovative and creative in their learning (p. 7)." It was not reported how the student-created "values" coincided or not with established learning standards or instructor expectations. The potential disparity in these valuesmay be worthy of investigation; alternatively, an emphasis on recognizing worth in student-created values may lead to further benefits or unique experiences for students and learning to occur.

K. Bartholomew et al. (2018a) research the potential correlation between ACJ and traditional rubric-based approaches to assessment. Further, they compared the ACJ and traditional grading results with the actual performance of the student designs. Their efforts "demonstrated a strong alignment between the ACJ-produced rankings and traditional rubric assessment methods (p. 31)." While this alignment suggests reliability in the results obtained through ACJ, neither ACJ nor the traditional grading approaches were significantly correlated with the actual performance of the design projects. The non-significant correlation suggests issues with validity in both assessment approaches if the desired outcome is to produce quality products. An increased understanding of the desired outcomes (i.e., quality products, evidence of improvement, learning goals) and students experience and effort towards achieving those outcomes will further illuminate these findings and the potential for additional effort and focus.

N. Seery et al. (2018) had university students complete several design assignments and then use ACJ to assess their peers for each assignment. ACJ was used to help track student ability levels over time. Students who were performing poorly initially, moved up in rank order over the course of the four assignments (p. 711). While this was true for those in the middle as well, top performers were the only ones who decreased in performance. The authors note the higher potential for initially-poor-performing students to increase—as opposed to those initial top performers—but also highlight the fact that as students progressed between assignments the entire body of students was changing. The authors noted that:

> there is little doubt that being exposed to superior work enhanced the performance of [low performing] students, it is speculated that the resulting goal settings and associated motivation resulting from the peer review enhanced their engagement and comprehension in subsequent assignments (p. 712).

While the authors call into consideration the potentially motivating experience of low performing students (i.e., being exposed to higher quality work), there is no discussion around the potentially de-motivating experience of top performing students; if they were all exposed to lower-quality peer work through ACJ could this possible explain why some of them performed worse on subsequent assignments? Efforts to explore the experiences of top-performing and low-performing student motivation, while engaged in ACJ assessments, would shed light on these possibilities and modifications which may improve the educational experience for all students.

R. Canty et al. (2019) had students (student teachers in education) create an activity and an assessment to go along with that activity. These deliverables were assembled into portfolios that were peer-judged using ACJ. Students were asked to make comments on the portfolios as they judged them. Based on the student survey responses following the experiment, students felt that using ACJ was helpful to their learning. Specifically,

students felt the experience broadened their understanding of the assignment and potential approaches. Student comments, left while completing judgments, were qualitatively analyzed and the results showed that the majority of comments focused on task or process level changes. The authors note that more thought-provoking comments (e.g., self-regulative level) were few and could likely be increased through an intervention or training in class. Investigating the potential for such an intervention to elevate student ACJ comments would strengthen the understanding of student experience and the potential for ACJ to be used as a learning tool and not just an assessment tool.

V. Buckley et al. (2020) had aspiring design and technology teachers use ACJ to evaluate design portfolios created by their peers. While completing ACJ comparisons no external criteria was provided, and the students were requested to leave comments explaining the criteria used in judgments for each comparison. A correlation between the contents of the student portfolios (e.g., how many images were included) and the ACJ outcome (i.e., parameter value) was conducted; the authors noted that (p. 11):

> nearly all quantitative variables describing the amount of content correlated significantly with performance, there is an indication that the amount of work at least aligned with perceptions of capability.

A second analysis was conducted to compare the approaches employed by students and their own judgment decisions (i.e., did their own designing style influence their comparative decisions such that they were inclines towards those that designed similarly to themselves?). The authors noted that their analysis showed—albeit without practical significance—that a portfolio was more likely to win when it was more different to the judges own portfolio than the comparison portfolio. A final analysis of student comments was conducted to identify the prevalence of varying rationales for judgment across student comments. They report that (p. 16): "the quality of the craft, alignment with the brief (in terms of conveying emotion), and quality of the portfolio were the most significant indicators of good performance." While they authors found that quality of craft, portfolio, and alignment with the brief were the most significant indicators of performance they also noted that, in the student's comparative decisions, quantity seems to be indicative of quality. This idea, which seems to contradict their other findings, is both intriguing and alarming; could the value of a portfolio be instantly raised simply through the addition of "filler" material? This idea warrants future investigation as it may undermine the trustworthiness of the ACJ approach altogether. Further, and potentially relatedly, the authors reported that no agreement statistics were computed for their qualitative analysis of student comments—this is an area for future efforts in line with ensuring sound findings and providing a second insight into their outcomes.

ACJ used with Multiple Approaches. These studies were labeled as "multiple approaches" because they used both formative and summative assessment, or an assessment method with other learning approaches such as peer-feedback or critique. Some studies also incorporated the use of surveys in either pre/post or post only fashion.

I. Potter et al. (2017) employed a mixed-methods approach with ComPAIR (an ACJ tool) with several different adaptations in each of several classes. The student survey results suggested that ComPAIR was a useful teaching tool to be used by teachers and students in various academic disciplines; something that was particularly true in ComPAIR's "capacity to strengthen students' abilities to self-assess the quality of their own answers to an assigned question and in fostering the ability to evaluate and provide feed-

back on others' work" (p. 111). Further, students rated the ACJ process of comparing peer answers and evaluating their own answers as a significant and important aspects of their learning—especially in Physics and English courses. However, students enrolled in Math courses did not perceive the benefits of ACJ similarly with much lower responses than the other subject areas. This is interesting to note as other studies with ACJ and Mathematics (Jones and Alcock 2012) have reported positive findings from students. Further investigation into the differences in student experience—based on subject area—is needed to increase the understanding of the potential for ACJ in classrooms. Additionally, understanding the nuances of ACJ-subject-area experiences may lead to further questions, improvements, and adjustments which lead to student learning.

J. Rhind et al. (2017) had university students use ACJ to judge student answers (from the previous year) and conducted a survey with students on their experience. Further, the authors investigated the student-produced ranking and the faculty scoring of the items. A significant correlation was found between the student ranking and the faculty marks as well as a weak positive correlation between the increase of time spent on the judgments and their own performance on the exam. While 78% of students agreed that they had learned from the ACJ process, slightly less than 50% of the students agreed the exercise was a good use of time. Future research efforts into why less than half the students perceived ACJ as a good use of time is needed; if students report learning from the process, and demonstrate high levels of correlation with instructor results, why don't they perceive the process as a good use of time? Additional qualitative efforts to understand this perception are important if ACJ is to be used as a learning tool for students as they are not likely to engage in a process/approach they do not perceive value in.

L. Barber (2018) assigned academic staff to use ACJ to grade/mark scripts that had been graded previously. ACJ marks given by the academic staff, though not well-correlated to previous grades, were "better than the original marks" obtained by earlier graders (p. 6). The author highlights the potential of ACJ for improving reliability in grading while also stressing the importance of selecting the judging criteria carefully. Additionally, university students peer-graded a mock examination question using ACJ and generally praised the approach for learning and were more inclined to this approach of peer-grading than assigning traditional marks (scores). Further investigation into the selection of judgment criteria, or the lack thereof (see Canty et al. 2017, 2019; Seery et al. 2016) and its impact on the selection process is needed. Judges selecting the better of two items, with different criteria for selection in mind, may arrive at the same conclusion, or not; regardless, an understanding of the impact of the criteria for selection is needed to further explore the ramifications of ACJ for assessment and learning.

M. Bartholomew et al. (2018b) used mixed methods approaches to gather various data across five studies of ACJ in design and technical education. The findings from each of the five studies demonstrated that ACJ had the potential to be an assessment tool for student competencies in educational settings. Implications for using ACJ as a learning tool in highly specialized fields are shared as well as future directions for research into the potential for learning these skills through comparison. The authors note difficulties in assessment as a result of uncertainty around the competencies desired, as well as the criteria for demonstrating competence. The authors recommend clearly identifying these before conducting an ACJ assessment of student work and highlight the need for additional research before widespread conclusions could be drawn.

Q. Demonacos et al. (2019) used ACJ as a peer-feedback tool at the end of a writing assignment. A survey was also used to gather data. Students and instructors demonstrated low levels of agreement (correlationally) over grades, however the authors noted

the quality of student and instructor qualitative feedback provided through ACJ. Interestingly, the authors note that the instructors and students provided feedback on very different aspects of the assignment. The differences in feedback between instructors and students is similar to a finding in Zhang (2019); an understanding of these differences, why they occur, and how the differences in feedback may influence subsequent student decisions, is unclear.

S. Zhang (2019) tasked several students with engaging in ACJ judgments of design projects. These judgments were also completed in a separate session by instructors; the comparison revealed that the correlation between ranks was not significant; further, students took significantly longer than instructors to complete the exercise. Specifically, Zhang cites students spending more time to describe their feelings and assess each design and a lack of comfort with specific design language as potential reasons for the significant increase in time spent. Zhang notes (p. 70):

> Students utilized more time evaluating the good and bad of each design and then deciding which one was better; however, instructors with more teaching experience appeared to use experience and instinct to make decisions more quickly and with less discussion.

Zhang notes that additional research into the role of context-specific language in the ACJ assessment experience is needed; understanding the potentially reciprocal role between understanding content, using appropriate language, and providing feedback may provide insight into new ways for using ACJ as a learning and assessment tool.

U. Strimel et al. (2020) utilized a mixed methods approach to gather both qualitative (student written comments and questionnaire) and quantitative data (ACJ data output). In this study ACJ was used as a formative tool (by students) and summative assessment approach (by students, instructors, and industry experts). The researchers investigated the potential for ACJ to be used as a learning tool as well as the correlation between summative ACJ results obtained from students, instructors, and industry experts. The authors found that involving students in ACJ for learning (i.e., before design work) allowed them to gain personal insights into the design process and improve their designing.

Further, the authors note that ACJ appeared to be an effective summative assessment approach for engineering design practices in education; however, there were stark differences in how students, instructors, and practicing engineers evaluated the design work and their ACJ-produced ranks. The authors note the need for additional research into the differences in group perceptions of quality as well as the implications of such differences. Specifically, the authors found a stronger correlation between the student-produced and industry-produced rank orders than the teacher-produced rank—understanding why this difference exists and how it may impact learning is needed to ensure effective learning and classroom expectations.

ACJ for Formative Assessment. These studies used ACJ explicitly as a formative assessment tool in an educational setting.

T. Bartholomew et al. (2019a) divided students in a large undergraduate course focused on design thinking into control and experimental groups. All students were given an assignment to develop point-of-view (POV) statements while designing; these POV statements focus on describing a need, user, and an insightful solution to a problem. Treatment groups students used ACJ to review POV statements collected from a previ-

ous year while control group students engaged in a teacher-led think-pair-share activity to prepare. All students completed POV statements; these were collected, anonymized, and then judged by instructors in an ACJ session. The treatment group POV statements were ranked significantly better than those produced by students in the control group (i.e., that did not complete the ACJ session). However, the authors cautioned that the instructors noted that none of the POV statements, from either group, were of high quality. Future research into the implications of a rank where items are ordered but do not demonstrate quality is necessary; if students are learning through ACJ, are they learning to produce quality work or simply to outperform their peers?

O. Bartholomew et al. (2019b) had a control and treatment group of graphic design students create design briefs and engaged the treatment group in ACJ to both provide and receive peer feedback. Additionally, all student work, from both groups, was graded using traditional approaches and a comparison of traditional versus ACJ methods was analyzed. Students results from those who used ACJ were not significantly better than their peers, however, their improvement levels in performance were significantly better when compared to their peers who did not. The authors noted that the benefits of ACJ included the exposure to new ideas through peer evaluation and the act of both providing and receiving feedback. However, the researchers also found that students did not trust peer feedback as much as that received from an instructor. This mistrust of peer feedback may prove to be a challenge for further ACJ implementation and warrants further investigation.

P. Bartholomew et al. (2019c) divided university graphic design students into three groups and completed varying amounts of ACJ sessions based on the section enrollment. Some students completed ACJ prior to engaging in design, others did ACJ in the midst of designing, while a third group completed ACJ at the conclusion of designing. All student work was assessed at the conclusion of the study and a survey reflection of student experiences with ACJ experience was also used. The authors noted no significant different in student achievement between groups. This finding, which contradicts the findings from other research (e.g., Bartholomew et al. 2019a) highlights the need for additional research into the potential for ACJ—a tool originally designed for assessment—to be used as a tool for learning. How often, when, and with whom ACJ should be used for learning are all questions for additional research.

ACJ implementation results

At the foundation of ACJ lies the idea that reliability of assessment results may be improved through a CJ process over more-traditional scoring approaches (Pollitt 2004, 2012; Thurstone 1927). In light of this idea, the majority of ACJ-related research includes a report of the reliability of the ACJ-produced results (e.g., rank order, parameter value); however, this is not always the case. As noted earlier, the "reliability" of ACJ-produced results has been debated (e.g., Bramley 2015; Pollitt 2015; Rangel-Smith and Lynch 2018) and the term "reliability" in ACJ-settings actually refers to a *JCC* (see Pollitt 2015) related to the judgments made by judges. Put differently, the reliability of the ACJ-produced results relates to the consistency with which judges select certain items as "better" than others.

Validity, as reported in ACJ research, most often refers to a comparison of ACJ-produced results with other forms of assessment and/or grading (i.e., traditional rubric-based grading or professional marking). In other scenarios the ACJ-produced rank was

compared with the actual functionality of designs or other commonly accepted standards as a "validity check." A synopsis of findings, related to both the reliability and validity of research around ACJ, is included in Table 3.

Conclusion

This synthesis was conducted with the intent of serving as a "launchpad" for those interested in ACJ in higher education settings, with key findings, efforts, and research synthesized for ease of review and identification of overarching findings and principles. This analysis focused on findings from studies of ACJ in higher education settings and, based on our synthesis of related research, we believe ACJ represents a valuable tool for a wide variety of content and contextual settings in higher education. As higher education assessment methods must adapt to the ever-increasing volatility, uncertainty, complexity, and ambiguity of our modern world, pedagogical practice and instructional design may end up requiring less "right" answer approaches to assessment and more "open-ended" methods. Therefore, ACJ may yet prove to be a vital learning and assessment tool in higher education. However, we caution readers that several areas of further inquiry were apparent in our review—these should be studied with intentionality before broadly accepting ACJ for educational purposes. Applications in both summative and formative assessment have demonstrated high levels of both reliability and validity as well as increases in student learning and achievement. Further, recent years have seen a marked increase in ACJ implementation in terms of location, quantity of research articles, and different content areas suggesting increased traction and interest surrounding the potential for increased use.

In addition to summarizing pertinent information from ACJ in higher education settings, this review exposes areas necessitating future research efforts. For example, the majority of ACJ research has been confined to a relatively small number of locations, content areas, and populations. Efforts in broadening the application, adoption, and investigation into ACJ would further strengthen the overall understanding on ACJ's potential in higher education. Questions around differences in results from various groups of judges (e.g., students, teachers, professionals) need to be investigated to illuminate the reasoning behind these differences. Further, research into the validity of ACJ results (i.e., does the "best" item in a rank actually represent something of "high quality"?) is needed before ACJ, as a tool for assessment, can be broadly implemented.

Significantly, while the majority of research in ACJ in higher education has emphasized summative applications, many recent studies have exposed the potential for using ACJ in formative settings—specifically as a learning tool for students. This recent shift towards utilizing ACJ in higher education formative settings represents a valuable area for additional research into the potential for positive classroom implementation. An effort into further investigating the potential for ACJ in higher education formative settings with specific emphasis on the method as a tool for student learning and the temporal intervals by which it is introduced would do much to advance our knowledge of the theoretical and practical boundaries and efficacy of ACJ. Additionally, these efforts coupled with further research into both the reliability and validity of ACJ may shed additional light into the benefits and challenges of this approach as well as the implications of increased adoption in higher education.

Table 3 Reliability and validity results across studies

ID	Author (year)	Validity (i.e., comparison with traditional grading or otherwise as a "validity check"). Results reported as presented in the corresponding research	Reliability level (JCC). Results reported as presented in the corresponding research
A	Seery, N., Lane, D., & Canty, D. (2011)	Spearman's Rho: −0.11 (student rank vs. weighted mean square score as judge)	r = 0.961
B	Jones, I. & Alcock, L. (2012)	Expert/Peer correlation: r = 0.628 Expert/Novice correlation: r = 0.546 Novice/Peer correlation: r = 0.666	Peer: r = 0.91 Novices: r = 0.99 Experts: r = 0.97
C	Seery, N., Canty, D., & Phelan, P. (2012)	Not reported	r = 0.955
D	Canty, D. (2012)	(Year 1) Pearson's Correlation of Rank 1 and Rank 2 parameter values: r = 0.695 (Year 2) Spearman's Correlation of student rank order position and judging misfit statistic: r = −0.162 (Year 3) Professional assessor and student assessor rank order correlations: (Y1) r = 0.917, (Y2) r = 0.849, (Y3) r = 0.88	(Year 1) Reliability of ACJ rank order: Rank 1 = Cronbach alpha α = 0.948; Rank 2 = Cronbach alpha α = 0.955 (Year 2) Reliability of ACJ rank order: Cronbach alpha α = 0.942 (Year 3) Reliability of ACJ rank order: Rank A to Rank B = α = 0.787 Rank A to Rank C = α = 0.702 Rank B to Rank C = α = 0.671
E	Rowsome, P., Seery, N., & Lane, D. (2013)	Not reported	r = 0.82
F	Seery, N., et al. (2016)	Correlation with traditional peer grading; r = 0.760 to r = 0.956	Interrater reliability of 0.961
G	Metzgar, M. (2016)	Compared to traditional grading; no data reported	Reliability of five assignments reported: r = 0.77; r = 0.81; r = 0.56; r = 0.76; r = 0.82
H	Canty, D., Seery, N., Hartell, E., & Doyle, A. (2017)	Not reported	r = 0.98
I	Potter, T. et al. (2017)	Not reported	Not reported
J	Rhind, S. M., et al. (2017)	Correlation with traditional grading; r = 0.690	r = 0.98
K	Bartholomew, S.R., Strimel, G.J., & Jackson, A. (2018)	Correlation with traditional grading; r = 0.79	r = 0.95
L	Barber, J. (2018)	Correlation with traditional grading; no data reported	r = 0.95 r = 0.94

Table 3 (continued)

ID	Author (year)	Validity (i.e., comparison with traditional grading or otherwise as a "validity check"). Results reported as presented in the corresponding research	Reliability level (JCC). Results reported as presented in the corresponding research
M	Bartholomew, S.R., Yoshikawa, E., & Connolly, P.E. (2018)	Not reported	Reliability of five domain area studies reported: r = 0.54 and 0.31(Transdisciplinary Studies in Technology); r = 0.94, 0.97, and 0.20 (Theatre); r = 0.84 (Industrial Design); r = 0.798, 0.719, −0.372 (Computer Graphics Technology); r = 0.56, 0.66, and 0.33 (Engineering Education)
N	Seery, N., et al. (2018)	Correlation with traditional grading; no data reported	Cronbach's Alpha reported for four assignments: α = 0.974; α = 0.973; α = 0.965; α = 0.971
O	Bartholomew, S. R., Zhang, L., Garcia Bravo, E. & Strimel, G. J. (2019b)	Correlation with traditional grading; r = −0.65	Reliability reported for two projects: Project 1, r = 0.78; Project 4, r = 0.85
P	Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J. (2019c)	A one-way ANOVA was used to investigate the impact of ACJ timing on student achievement using parameter value means. Beginning group mean: −0.216; Middle group mean: −0.016; Beginning and middle group mean: 0.142. There was no significant difference among groups	Not reported
Q	Demonacos, C., Ellis, S., & Barber, J. (2019)	Made qualitative correlations with traditional grading; no data reported	R = 0.55
R	Canty, D., Buckley, J., & Seery, N. (2019)	Not reported	Cronbach alpha of α = 0.68
S	Zhang, L. (2019)	Spearman's Rho for correlations between student and instructor ranks: r = 0.564 Spearman's Rho for correlations between student and instructor parameter values: r = 0.465	Student session: r = 0.83 Instructor session: r = 0.80
T	Bartholomew, S.R., Mentzer, N., & Jones, M. (2019)	Correlation between student and instructor ranks:r = 0.57	Not reported

Table 3 (continued)

ID	Author (year)	Validity (i.e., comparison with traditional grading or otherwise as a "validity check"). Results reported as presented in the corresponding research	Reliability level (JCC). Results reported as presented in the corresponding research
U	Strimel, G.J., Bartholomew, S.R., Purzer, S., Yoshikawa, E., & Zhang, L. (2020)	Correlations reported for ranks produced by students and teachers: r = 0.564; ranks produced by students and practicing engineers: r = 0.66; and for ranks produced by teachers and practicing engineers: r = 0.33	Not reported
V	Buckley, J., Canty, D., Seery, N.	Auditing by module leaders through individually grading each piece of work for use and monitoring of misfit statistics. Qualitative analysis and coding to describe the amount of work, independent of quality, which was done (e.g. the number of sketches presented)	r = 0.974

References

Akister, J., Bannon, A., & Mullender-Lock, H. (2000). Poster presentations in social work education assessment: A case study. *Innovations in Education and Training International, 37*(3), 229–233.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*(2), 114–129.

Barber, J. (2018). Five go marking an exam question: The use of adaptive comparative judgement to manage subjective bias. *Practitioner Research in Higher Education, 11*(1), 94–100.

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

Bartholomew, S. R., Mentzer, N., & Jones, M. (2019a). Learning by evaluating (LbE). In *Mississippi Valley Technology Education Conference*. Nashville, TN.

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

Bartholomew, S. R., Yoshikawa, E., & Connolly, P. E. (2018b). Exploring the potential for identifying student competencies in design education through adaptive comparative judgment. In *PATT35 Athlone Institute of Technology, Athlone, Ireland 18–21 June, 2018* (pp. 187–194).

Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J. (2019b). A tool for formative assessment and learning in a graphics design course: Adaptive comparative judgement. *The Design Journal, 22*(1), 73–95.

Bartholomew, S., Zhang, L., Bravo, E. G., & Strimel, G. J. (2019c). First, last, elsewhere… positioning adaptive comparative judgment in the design learning experience. In *PATT 37 University of Malta, Malta, June 2–6, 2019* (pp. 65–74).

Bell, S. (2010). Project-based learning for the 21st century: Skills for the future. *The Clearing House, 83*(2), 39–43.

Benton, T., & Gallagher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment Publication, 26*, 22–28.

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.

Boud, D., & Feletti, G. (2013). *The challenge of problem-based learning*. London: Routledge.

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

Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43–58.

Buckley, J., Canty, D., & Seery, N. (2020). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies.* https://doi.org/10.1080/03323315.2020.1814838.

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

Canty, D., Buckley, J., & Seery, N. (2019). Inducting ITE students in assessment practices through the use of comparative judgment. In *PATT 37 University of Malta, Malta, June 2–6, 2019* (pp. 117–124).

Canty, D., Seery, N., Hartell, E., & Doyle, A. (2017). Integrating peer assessment in technology education through adaptive comparative judgment. In *PATT34 Technology & Engineering Education–Fostering the Creativity of Youth Around the Globe, Millersville University, Pennsylvania, USA* (pp. 10–14).

Demonacos, C., Ellis, S., & Barber, J. (2019). Student peer assessment using adaptive comparative judgment: Grading accuracy versus quality of feedback. *Practitioner Research in Higher Education, 12*(1), 50–59.

Dobson, S. (2006). The assessment of student PowerPoint presentations—Attempting the impossible? *Assessment & Evaluation in Higher Education, 31*(1), 109–119.

Duran, M., & Dökme, İ. (2016). The effect of the inquiry-based learning approach on student's critical-thinking skills. *Eurasia Journal of Mathematics, Science & Technology Education, 12*(12), 2887–2908.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91–108.

Hannafin, M. J., Hall, C., Land, S., & Hill, J. (1994). Learning in open-ended environments: Assumptions, methods, and implications. *Educational Technology, 34*(8), 48–55.

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 & Technology Education: An International Journal, 12*(2), 66–76.

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.

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

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.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718.

Martin-Martin, A., Orduña-Malea, E., Harzing, A. W., & López-Cózar, E. D. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics, 11*(1), 152–163.

Metzgar, M. (2016). Using adaptive comparative judgement to assess student work in an MBA course. *International Journal for Infonomics, 9*(3), 1217–1219.

Mills, J. E., & Treagust, D. F. (2003). Engineering education—Is problem-based or project-based learning the answer. *Australasian Journal of Engineering Education, 3*(2), 2–16.

Moskal, B. M., Leydens, J. A., & Pavelich, M. J. (2002). Validity, reliability and the assessment of engineering education. *Journal of Engineering Education, 91*(3), 351–354.

Munroe, L. (2015). The open-ended approach framework. *European Journal of Educational Research, 4*(3), 97–104.

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). Rotterdam: Sense Publishers.

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri, 55*(4), 170–180.

Pollitt, A. (2004). *Let's stop marking exams*. Retrieved July 23, 2018, from http://www.cambridge assessment.org.uk/images/109719-let-s-stop-marking-exams.pdf.

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

Pollitt, A. (2015). *On 'reliability' bias in ACJ*. Cambridge Exam Research. Retrieved February 2, 2018, from https://www.researchgate.net/publication/283318012_On_'Reliability'_bias_in_ACJ.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Studies in Language Testing, 3*, 74–91.

Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I. (2017). ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry, 5*(2), 89–113.

Purzer, S., Fila, N., & Nataraja, K. (2016). Evaluation of current assessment methods in engineering entrepreneurship education. *Advances in Engineering Education, 5*(1), n1.

Rangel-Smith, C., & Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment. In *PATT35 Athlone Institute of Technology, Athlone, Ireland 18–21 June, 2018* (pp. 378–387).

Rhind, S. M., Hughes, K. J., Yool, D., Shaw, D., Kerr, W., & Reed, N. (2017). Adaptive comparative judgment: A tool to support students' assessment literacy. *Journal of Veterinary Medical Education, 44*(4), 686–691.

Rowsome, P., Seery, N., & Lane, D. (2013). *The development of pre-service design educator's capacity to make professional judgments on design capability using adaptive comparative judgment*. American Society for Engineering Education.

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2018). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education, 29*, 701–715. https://doi.org/10.1007/s10798-018-9468-x.

Seery, N., Buckley, J., Doyle, A., & Canty, D. (2016). The validity and reliability of adaptive comparative judgements in the assessment of graphical capability. In *Proceedings of the 71st Mid-Year Conference of the Engineering Design Graphics Division* (pp. 104–109).

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., Lane, D., & Canty, D. (2011). Exploring the value of democratic assessment in design based activities of graphical education. In *118th Annual American Society of Engineering Education Conference*. Vancouver, BC: American Society for Engineering Education.

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

Stevens, D. D., & Levi, A. J. (2013). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VA: Stylus Publishing, LLC.

Strimel, G. J., Bartholomew, S. R., Purzer, S., Yoshikawa, E., & Zhang, L. (2020). Informing engineering design through adaptive comparative judgment. *European Journal of Engineering Education*. https://doi.org/10.1080/03043797.2020.1718614.

Thomas, J. W. (2000). *A review of research on project-based learning*. Autodesk Foundation. Retrieved from https://www.asec.purdue.edu/lct/HBCU/documents/AReviewofResearchofProject-BasedLearning.pdf.

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

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562.

Viseu, F., & Oliveira, I. B. (2017). Open-ended tasks in the promotion of classroom communication in mathematics. *International Electronic Journal of Elementary Education, 4*(2), 287–300.

Zhang, L. (2019). *Investigating differences in formative critiquing between instructors and students in graphic design*. Doctoral dissertation, Purdue University Libraries.

Zientek, L. R., Werner, J. M., Campuzano, M. V., & Nimon, K. (2018). The use of Google Scholar for research and research dissemination. *New Horizons in Adult Education and Human Resource Development, 30*(1), 39–46.