An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education

Jeffrey Buckley[a,b]*, Donal Canty[c] and Niall Seery[d]

[a]Faculty of Engineering and Informatics, Athlone Institute of Technology, Athlone, Ireland

[b]Department of Learning, KTH Royal Institute of Technology, Stockholm, Sweden

[c]School of Education, University of Limerick, Limerick, Ireland

[d]Office of the President, Athlone Institute of Technology, Athlone, Ireland

*corresponding author email address: jbuckley@ait.ie

An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education

The use of design assignments for teaching, learning, and assessment is considered a signature of technology education. However, there are difficulties in the valid and reliable assessment of features of quality within designerly outputs. In light of recent educational reforms in Ireland, which see the introduction of classroom based assessments centring on design in the technology subjects, it is paramount that the implementation of design assessment is critically considered.

An exploratory study was conducted with a 1st year cohort of initial technology teacher education students (N = 126) which involved them completing a design assignment and subsequent assessment process through the use of adaptive comparative judgement (ACJ). In considering the use of ACJ as a potential tool for design assessment at post-primary level, data analysis focused on criteria used for assessment. Results indicate that quantitative variables, i.e., the amount of work done, can significantly predict performance ($R^2 = .333$, $p <.001$), however qualitative findings suggest that quantity may simply align with quality. Further results illustrate a significant yet practically meaningless bias may exist in the judgement of work through ACJ ($\varphi = .082$, $p <.01$) and that there was need to use varying criteria in the assessment of design outputs.

Keywords: adaptive comparative judgement; assessment; design; educational reform; technology education

Introduction

Technology education in Ireland at post-primary level consists of the four subjects of Wood Technology, Applied Technology, Engineering, and Graphics at Junior Cycle (lower post-primary education, students aged ≈ 12-15), and as the four subjects of Construction Studies, Technology, Engineering, and Design and Communication Graphics at Senior Cycle (upper post-primary education, students aged ≈ 16-18). Traditionally, Irish technology education was considered to be craft orientated with possible movement towards a design approach (Carty and Phelan 2006), however, contemporary technology education in Ireland, as a result of the recent reforms at Junior and Senior Cycle, illustrate that philosophical shift towards embracing design based education.

In the current Junior Cycle reform, the assessment mechanism is of particular pertinence for the technology subjects due to the prevalence of open-ended project work which largely takes the format of design assignments. In the new Wood Technology and Applied Technology subjects the emphasis is now placed on design as an iterative process (NCCA 2018c, 2018a), design in the Engineering subject is now being framed as cyclical problem solving (NCCA 2018b), and design in the Graphics subject is conceptual with the ‘make’ element taking the form of a computer aided design (CAD) model much like the Senior Cycle Design and Communication Subject (NCCA 2019).

In addition to the introduction and reframing of design in the Junior Cycle technology subject’s project work, the new specifications see the introduction of classroom-based assessments (CBA’s) presenting two additional critical milestones for formative feedback based on open-ended projects in each subject. Therefore, there are now three critical, national level, design-based projects in each of the Junior Cycle technology subjects, two CBA’s and a final project. These require valid and reliable assessment mechanisms for formative and summative purposes. As there are many difficulties associated with the assessment of such open-ended project work in technology education (Kimbell 2007), there is a need to consider mechanisms for the enactment of design assessment to ensure an equitable and fair education system for all students.

Assessing design in technology education

Sadler (2009) describes two significant implications for the validity of assessment of open-ended tasks using criterion-referenced assessment such as what is done in technology education in Ireland. The first is that the sum of the criterion-referenced scores may not align with the holistic professional opinion of the assessor, and the second is that a rubric may not adequately account for the idiosyncrasies inherent in a student's work such that a student may do something exceptional yet unexpected. These issues are heightened in technology subjects as students can demonstrate competency or excellence in dramatically different ways (Kimbell 2007) with outcomes of design solutions often involving more variables than can be incorporated into assessment criteria (Williams 2000). In addition to this, the validity of assessing the design aspect to project work in the Irish technology subjects is questionable, as it is with design work in many other areas. There is a craft element which is objectively visible through the artefacts which are produced by students, however the design journey is represented through a diary-style portfolio. While design in the Irish technology subjects has theoretically always placed emphasis on the process as well as the product of learning through design as evidenced by the inclusion of a portfolio, the inclusion of the portfolio alone does not imply alignment of the intended and assessed curricula (Kurz et al. 2010). For example, while the aim of a portfolio may be to describe the process of design and of students learning, it may be assessed as a product in and of itself if assessors are biased towards the quality of outputs within the portfolio. Furthermore, if students treat design portfolios as products rather than as documentation of their design journeys in attempts to conform to assessment criteria and align outputs to specific criteria they bring into question the validity of the use of portfolios as an assessment instrument (Seery, Canty, and Phelan 2012).

In light of this, the problem with assessing design related outcomes in technology education is that the assessment is “trying to measure evidence of thinking while encouraging diversity within a system predicated on standardisation and weighted criteria" (Seery, Canty, and Phelan 2012, 208). However, despite the difficulties with assessing design, there are many potential solutions. Perhaps the most widely researched method to addressing this problem, at least within technology education, is the use of adaptive comparative judgement (ACJ). Pioneered through project e-scape for design related outputs in technology education (Kimbell et al. 2009, 2005, 2007), ACJ involves a group of assessors making judgements on a range of projects based on Thurstone's (1927) Law of Comparative Judgement and has been found to be a highly reliable method to assess design related outputs (Bartholomew and Yoshikawa-Ruesch 2018).

Adaptive comparative judgement

The method of adaptive comparative judgement

A number of studies present comprehensive explanations of the ACJ method. Pollitt (2012a, 2012b) describes the process in detail with specific emphasis on the underpinning theory and mathematics, while Seery, Buckley, Delahunty and Canty (2018) provide a detailed account of the use of ACJ in technology education practice. These studies, in conjunction with the wide variety of applications of ACJ across multiple different contexts, described in a systematic review by Bartholomew and Yoshikawa-Ruesch (2018), suggest that a complete account of the process is not required so as to avoid repetition within the literature. However, the majority of previous studies focus on ACJ as an assessment tool, and while this is considered in the current study, much of the focus here is placed on the qualitative commentary that ACJ facilitates the collection of. For this reason, a brief account of the ACJ method will be provided to contextualise the collection of qualitative data.

The ACJ method involves three components, (1) an assortment of digitised pieces of work which are the subject of the assessment process, (2) a cohort of people who use the ACJ method to assess the work, and (3) the ACJ system which is a web-based system that controls the assessment process. For coherency in this paper and with the pertinent literature, the pieces of work subject to assessment will be described as ‘portfolios’ and the assessors will be described as ‘judges’. Once all portfolios are uploaded to the ACJ system, the process begins by presenting a judge with two of them. The judge then makes a determination on which of the two portfolios is better, and this decision can be based on a holistic judgement or external itemised criteria depending on the agenda of the assessment. Once a decision is made, judges are given an opportunity to provide two types of commentary. The first is commentary about each portfolio which can serve as a feedback function for students. The second is commentary explaining why they chose a particular portfolio as being better than the other. This can be used in multiple ways, such as an audit mechanism to qualify spurious judgments and also to gain insight into the qualities of work considered to be indicative of capability. This process then progresses with the judge being shown a new pair of portfolios to make a judgement on and to provide commentary for, and is repeated multiple times with multiple judges until the required number of judgements are made. The outputs of this process include a rank order from better to worse of all portfolios with relative distances in quality denoted by parameter values (*z*-scores), misfit statistics denoting both judge and portfolio outliers, and commentary provided by judges on each portfolio and on their decision processes. It is important to clarify that the rank order which is produced is relative and does not provide any denotation of quality other than relative performance. In practice, the highest ranking portfolio may not necessarily be high quality work and the lowest ranking portfolio may not be poor quality. They were simply judged to be the highest and lowest performing portfolios within the sample by that particular group of judges. The transposition of the rank to alternative descriptors such as grades requires an additional step and likely expert evaluation.

Benefits of comparative judgement

There are a variety of terms used to describe the fundamental process of comparative judgement in relation to assessment. These include, for example, comparative judgement, pairwise comparison, paired comparison, pairwise judgements, and comparative pairs. These terms are all synonymous, with the only exception being adaptive comparative judgement, which describes a difference in the sorting algorithm (Pollitt 2012b) but not in the experience of the judge. As such, all of these forms of comparative assessment share the same educational benefits. Arguably the largest benefit to the use of ACJ is the high level of reliably in the assessment of open-ended outputs which is regularly observed in studies which have used it. Within technology education, studies utilising ACJ consistently report interrater reliability values of >.93 (Kimbell 2012; Seery, Canty, and Phelan 2012; Bartholomew et al. 2017; Bartholomew, Strimel, and Jackson 2018; Seery et al. 2018) with reliability scores of >.80 frequently found in other subject areas (Pollitt 2012b; Jones, Swan, and Pollitt 2015; Newhouse 2014; Steedle and Ferrara 2016). This is achieved by changing the assessment question from asking assessors to assign specific marks to student outputs based on often ill-defined criteria to asking them to make a professional holistic judgement on quality between two portfolios, a typically easier assessment question to answer. Additionally, by including a variety of judges in the process, biases held by judges can theoretically be mitigated and therefore the validity of the assessment process can be increased. However, if the entire cohort of judges holds similar biases such mitigation cannot occur. In such a case questions of validity shift from a biased judge to a potentially biased task or general bias towards the nature of the activity. The addition of misfit statistics further adds to the validity of the use of ACJ, as if a portfolio or judge is an outlier, i.e. if there is inconsistency with judgments associated with a particular portfolio or which are made by a specific judge, these are visible and related qualitative commentary can be used as an audit mechanism. Another regularly purported benefit of ACJ is that the students can act as judges (e.g., Canty, Seery, and Phelan 2012; Seery et al. 2018, 2012), thus integrating them into the assessment process which is viewed as a valuable learning opportunity (Nicol and Macfarlane-Dick 2006; Sadler 2009).

A critical aspect of the adoption of ACJ is the time commitment taken to assess work. There is insufficient data to make an adjudication on this, as some studies suggest ACJ is quicker than traditional grading methods (e.g., Newhouse 2014; Steedle and Ferrara 2016), others suggest that traditional grading takes less time (Bartholomew and Yoshikawa-Ruesch 2018), while others suggest both approaches require a comparable time investment (Coertjens et al. 2017). However, it is unlikely that debating the time cost of the use of ACJ will ever lead to a simple answer due to the multitude of variables impacting the decision such as judge expertise, the format of the traditional rubric, the nature of the portfolios being assessed, the subject area, and the number of judges involved in the process. Instead, it is perhaps more appropriate to use existing evidence as a guide, however the decision should weigh up the benefits of potentially increased reliability and validity and the opportunity to integrate students into the assessment process with the cost and logistical implications. Finally, in relation to the adoption of ACJ as a tool for national assessment, logistically judges can be based anywhere as evidenced by Bartholomew, Yoshikawa, Hartell and Strimel (2019) where judges from Ireland, the UK, Sweden and the USA judged student work created in the

USA. As judges, likely to be teachers in a national context, adjudicate as a cohort, they are given more freedom in making professional judgements, and their decisions form a collective consensus such that no individual judge would be responsible for the grades assigned to individual students, a process that can be implemented in multiple ways by transforming the parameter values to percentages.

Study purpose

In viewing ACJ as a potentially viable method of assessment for design activities in Irish technology education, there is sufficient evidence to indicate that it could work in practice. However, pragmatic and logistical questions would need to be considered based on specific cases of implementation. A more critical question which must be considered is the need for such a mechanism. Conventional rubrics have traditionally been used and the adoption of an ACJ style approach would see a substantial paradigm shift in terms of the assessment process. In considering the need for an ACJ style approach to the assessment of design related outputs, the decision making process of judges needs to be explored with respect to the open-ended outputs typical of assignments in technology subjects, and therefore this study is centred on the research question of what variables, both quantitative and qualitative, do judges consider when making holistic comparative assessments of design related outputs when presented in an ACJ style environment.

Method

Approach and participants

To investigate the criteria used by judges when making comparative judgements, a cohort of 1st Year undergraduate students (N = 126) studying on an Initial Technology

Teacher Education (ITTE) programme in Ireland engaged with a design assignment and subsequent ACJ assessment of the work wherein they acted as the judges. Similar to previous work, the students who engaged with the task acted as judges as they were deemed best placed to empathise with the skillset needed to complete the assignment, and as ITTE students they had an insight into viewing such work through an assessment lens (Seery, Canty, and Phelan 2012; Seery et al. 2018). The module leaders did audit this process by individually grading each piece of work for use within the module and by monitoring misfit statistics, however the focus of this study is on the judgements used in assessment decision making which led to rank and not on the rank itself. In order to examine the criteria used by the students to make these judgements, they were requested to make qualitative commentary regarding their decisions during the ACJ process. Additionally, each portfolio was analysed and coded from a quantitative perspective to describe the amount of work, independent of quality, which was done, e.g. the number of sketches presented.

Design of instruments

The creation of design briefs in technology education is a complex task. Much evidence illustrates the existence of a context effect (e.g., Kimbell et al., 2004) whereby if "there was a personal context girls performed better, boys performed better when there was an industry context, and an environmental context proved to be more gender neutral" (Seery et al. 2019, 167). Therefore, a design task was created based on the work of (Seery, Canty, and Phelan 2012) which allowed for students to determine their own context. "Students were required to make an A4 framed pictorial scene with the composition of the scene being of the students own choosing, but portraying a dominant feeling or emotion. In addition, students were required to complete a second artefact.

They were challenged to design and make a flower (without facial expression) to express or reflect the emotion or feeling conveyed in their pictorial scene" (Seery, Canty, and Phelan 2012, 210). These projects were completed as part of a process technology model wherein the students were initially taught basic decorative processing techniques with wood (carving, laminating, and marquetry) and metal (planishing, scrollwork, enamelling, etching, repousse, mottling, and spinning). The project then was an open-ended and ill-defined opportunity for students to design and manufacture artefacts to both develop and evidence their learning. There were no restrictions on students other than how the task was previously described. Students did not have to present competency with any specific processing techniques, they had to determine their own constructs of capability within technology education and then use the design task as a conduit to present evidence of that.

In conjunction with the made artefacts, students were required to create a digital portfolio which captured their design journey. The portfolio consisted of a blank web space where students could create sequential panes (placeholders for files). A title could be added to each pane, and it could be filled with an unlimited amount of media files (images, audio files and videos) as well as text. Students could present their work in any order. There were no restrictions on the content of the portfolio. However, students were asked to apply tags to their work to describe if individual items represent them either having, growing, or proving an idea. Students could choose not to apply tags which may have been a deliberate decision if, for example, they were solely presenting their progress to date and they could add an appropriately descriptive title, but they were restricted to the three tags of having, growing and proving in line with Kimbell et al. (2004). In practice, the tagging process was implemented as a tool to provide opportunity for the students to reflect and analyse their thinking at that particular time in the learning process which also helped give visibility to both themselves and the module leaders on how they were progressing and whether or not they may be concentrating or fixating on a particular phase of the design learning journey. Therefore, a finished portfolio was in essence an electronic repository filled with a number of panes containing collections of media files and text, with each pane having a title and displaying an indication of the relative extents that it represented the student having, growing or proving ideas.

Implementation

The students completed the design assignment across two modules in a single 15-week semester (12 teaching weeks). The assignment formed a component of both modules assessment mechanisms providing motivation for the students to invest effort. Subsequent to the completion of the assignment, all students participated in an ACJ session where they made comparative judgements on the completed assignments. This aspect was not considered as part of the modules' summative assessment mechanism. Of the 126 students, 123 made 9 comparative judgements and 3 made 8 comparative judgements with no time limit. No external criteria were applied to the judgements, instead they were made holistically on criteria determined by the students individually. The only guidance given was that a judgement should be made based on the evidence of capability demonstrated through the portfolios. Students were requested to leave a comment at the end of each of their judgements explaining the criteria they used to make their decisions. This was completed through the ACJ interface so there could be alignment between the portfolios, the judges, and the comments.

Treatment of data

The data was analysed in three ways. First, the contents of each portfolio (N = 126) were objectively examined from a quantitative perspective. This included counting the total number of media files included with breakdowns for images, audio files, and video files, the average number of media files per pane with breakdowns for images, audio files and video files, and the percentage of the portfolios that the students denoted as themselves having, growing, or proving ideas. The output of the ACJ process is a rank order of the portfolios from better to worse, with relative distances displayed through parameter values. Correlations were examined to determine if there were significant relationships between the quantity of work presented and the performance of the portfolio as determined by its parameter value. This was followed by conducting a stepwise regression analysis to examine which quantitative variables had a predictive capacity for the students' performance, and how much variance could be explained.

The second aspect to the data analysis was to examine whether the students were biased in their judgments based on the similarity between their approach to the assignment and the portfolios they were making a decision on. This was done by comparing the percentage of having, growing and proving tags in the judge's portfolio with the percentages of those tags in the portfolios being judged through a chi-square test of independence. While this has a limitation in that it doesn't necessarily represent the nuances of how the students worked, in this study it is considered as a way of checking whether judges' biases for general ways of working could be mitigated through the ACJ process.

The final aspect of the data analysis involved coding and examining the frequencies of various rationales given by the students for their judgements. All three authors examined the comments provided by the students. The first author inductively coded each of the comments. This process included initially summarising the comments to qualitatively remove duplicates. For example, the comments "Project A was just better made" and "The project in portfolio B was manufactured better than the one in portfolio A" were coded as "Work quality" and noted that this was in reference to the quality of the craft. There was a reductionist objective but there was no upper limit on the number of codes. Importantly, a comment could receive multiple codes if more than one rationale was given for the decision. Each code with one (where only one instance was recorded) or two examples was then reviewed by the second author and an iterative process commenced until consensus was achieved as to what codes existed and were unique. This process resulted in a series of individual codes which were grouped into sub-categories which themselves were grouped into categories. This was then followed by both researchers individually coding each of the comments using the agreed upon codebook independent of the initial inductive process. Both researchers compared their coding of the data and any discrepancies were mutually reviewed. Finally, the last author reviewed the final coding and collective discourse ensued until authors agreed on the final codes. It is important to note that, in alignment with an interpretive epistemology, no agreement statistics were computed for this process, and findings should be considered accordingly. A correlation between the amount of criteria given by students across their judgements and their own performance was examined, and the frequencies for codes across all judgements are presented as *z*-scores.

Results

Quantitative predictors of performance

The reliability of the ACJ session was α = .974 indicating a very high level of consensus within the cohort as to the rankings of each portfolio and their parameter values. The mean time for the student's judgements in producing this rank was 10.108 minutes (standard deviation 10.701 minutes). However, a Shapiro-Wilk test indicated that this data was not normally distributed (*W* = .741, $p < 0.001$) and was skewed (skewness = 2.290) by a number of long judgements (max = 59.750 minutes) and therefore the median (6.683 minutes) and median absolute deviation (5.980 minutes) are more appropriate measures of central tendency and variance. In terms of interpreting this data, some of the variance is likely attributable to differences in the time students made their judgements. While the initial ACJ rounds which involve a rough sort through a Swiss tournament could see portfolios from both ends of the rank comparatively assessed, students making judgements in later rounds when the rank is largely qualified would be deciding between portfolios deemed to be closer in terms of relative quality. Beyond this, there is little that can be inferred from the amount of time spent making judgements as longer judgements don't necessarily translate to depth or criticality of thinking and it is possible that a person with a sophisticated, or at least steadfast, construct of capability could make similar comparisons in a fraction of the time.

Pearson's correlation coefficients were initially computed between each of the quantitative characteristics of the portfolios. While the full correlation matrix is shown in Table 1, only relationships between the different variables and the parameter values are of interest in this study. The correlation matrix includes portfolio rank position as a variable, however is not relevant to this study. Its inclusion is due to other studies often considering rank positions and not parameter values as the primary dependent variable (e.g., Bartholomew, Strimel, and Yoshikawa 2018) so in the event of future meta-analyses it is important to present both. A number of statistically significant correlations were observed indicating relationships between performance and the total number of panes ($r = .347$, $p < .001$), the total number of media files ($r = .506$, $p < .001$), the average number of media files per pane ($r = .473$, $p < .001$), the number of images taken from online sources ($r = .329$, $p < .001$), the total number of photographs ($r = .401$, $p <$

.001), the total number of student made images (photographs and sketches) (*r* = .411, *p* < .001), the total number of images (*r* = .484, *p* < .001), the average number of images per pane (*r* = .448, *p* < .001), the total number of student made videos (*r* = .377, *p* < .001), the total number of videos embedded from online resources (*r* = .250, *p* = .005), the total number of videos (*r* = .450, *p* < .001), the average number of videos per pane (*r* = .408, *p* < .001), and the total number of audio files (*r* = .182, *p* = .042). However, there were 18 correlations of interest examined so it is important to interpret these based on an adjusted alpha level (0.05/18) of 0.003. In this case, all correlations except for the relationships between performance and the total number of videos embedded from online resources (*r* = .250, *p* = .005) and the total number of audio files (*r* = .182, *p* = .042) can be interpreted as statistically significant. As 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, if indeed it was not a perceived descriptor of it. No significant correlations were observed between the percentages of having, growing, and proving ideas tags and performance, suggesting that the nature of the students' work in this regard did not relate to performance.

Table 1. Pearson's r correlation coefficients between the quantitative portfolio characteristics (N = 126).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Parameter value	-																		
2. Rank position	-.978**	-																	
3. No. of panes	.347**	-.330**	-																
4. No. of media files	.506**	-.487**	.525**	-															
5. Avg. media per pane	.473**	-.463**	.080	.852**	-														
6. No. of images (online)	.329**	-.328**	.240**	.509**	.445**	-													
7. No. of images (photos)	.401**	-.373**	.499**	.908**	.750**	.147	-												
8. No. of images (sketches)	.093	-.137	-.017	.161	.239**	.009	.000	-											
9. No. of images (student made)	.411**	-.391**	.489**	.922**	.780**	.147	.985**	.173	-										
10. No of images (total)	.484**	-.466**	.517**	.996**	.848**	.513**	.911**	.154	.924**	-									
11. Avg. images per pane	.448**	-.438**	.071	.845**	.994**	.453**	.751**	.228*	.779**	.851**	-								
12. No. of videos (student made)	.377**	-.320**	.181*	.196*	.164	-.088	.222*	-.069	.207*	.145	.107	-							
13. No. of videos (online)	.250**	-.264**	.201*	.329**	.282**	.389**	.148	.173	.176*	.303**	.253**	.011	-						
14. No. of videos (total)	.450**	-.411**	.261**	.344**	.291**	.142	.266**	.039	.268**	.288**	.229*	.833**	.562**	-					
15. Avg. videos per pane	.408**	-.371**	.111	.289**	.314**	.095	.224*	.058	.230**	.237**	.251**	.802**	.520**	.951**	-				
16. No. of audio files (total)	.182*	-.204*	.151	.247**	.215*	.059	.156	.151	.180*	.179*	.128	.126	.069	.142	.111	-			
17. Avg. audio files per pane	.151	-.175*	.044	.177*	.209*	.000	.102	.168	.129	.112	.120	.089	.041	.096	.078	.948**	-		
18. Having ideas %	.140	-.123	.030	.097	.076	.167	.041	-.080	.027	.092	.070	.048	.051	.069	.012	.059	.087	-	
19. Growing ideas %	.099	-.114	.179	.134	.036	.057	.126	.051	.132	.138	.045	-.075	-.096	-.116	-.127	.089	.014	-.141	-
20. Proving ideas %	-.181	.181	-.164	-.177	-.085	-.168	-.130	.017	-.125	-.177	-.087	.024	.039	.042	.092	-.113	-.075	-.619**	-.690**

Note. ** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).

A stepwise multiple linear regression was conducted with 10 independent variables and with performance as denoted by the portfolios parameter values as the dependent variable. A number of variables from the correlation matrix were excluded from being considered for the model as they were functions of other variables. For example, the ‘No. of images (total)’ variable is a function of the variables describing the total numbers of online images, photos and sketches so it wasn’t included. The final model (Table 2) was statistically significant (F(3,100) = 16.657, $p < .001$) and explained 33.3% of the variance in performance. Based on the regression model, the total number of videos students made positively influenced performance (β = .378), as did the total number of online images included (β = .320) and the total number of photos included (β = .246), so could be considered predictive of performance.

Table 2. Stepwise multiple linear regression with portfolio parameter values as the dependent variable.

Independent variables	ΔR^2	β	ΔF	*df*
Step 1				
No. of videos (student made)	.153	.378	18.394*	102
Step 2				
No. of images (online)	.123	.320	17.164*	101
Step 3				
No. of images (photos)	.057	.246	8.601*	100
Full model statistics				
Total R^2	.333			
Total R	.577			

Note. * $p < .01$. Independent variables included = No. of panes, No. of images (online), No. of images (photos), No. of images (sketches), No. of videos (student made), No. of videos (online), No. of audio files (total), Having ideas %, Growing ideas %, and Proving ideas %.

Analysis of tags

The results of each comparison from the ACJ session were analysed to examine the potential relationship between winning a judgement and a portfolios similarity to the judge's portfolio as denoted by the percentage of having, growing, and proving ideas tags. Similarity was determined by calculating the sum of the absolute percentage differences between the having, growing, and proving tags in a judges own portfolio and the portfolios being assessed. The total number of comparisons in the session was 918. Untagged portfolios were excluded from this analysis and therefore the total number of valid comparisons was 623.

Out of the 623 comparisons, 45.907% were won by the portfolio which was more similar to the judges own portfolio. A chi-square test of independence was performed to further examine the relationship between winning a judgement and a portfolios similarity to the judge's portfolio. A statistically significant relationship with a small effect size was found between these variables, χ^2 (1, N = 623) = 8.350, p <.01, φ = .082. These results suggest that a portfolio was more likely to win when it was more different to the judges own portfolio however the effect size was not of practical significance. Finally, an independent samples t-test was performed to examine the average magnitude of difference between winning and losing portfolios and a judges own portfolio. A statistically significant difference was not found between the average difference of winning portfolios (M = 39.925, SD = 22.329) and the average difference of losing portfolios (M = 37.980, SD = 20.337), t(1233.297) = 1608, p = .108.

Analysis of Judgement Criteria

A Pearson's correlation was examined between the parameter values of the participants' portfolios and the amount of criteria they referenced when rationalising their decisions.

A statistically significant weak correlation was found between the total amount of criteria given across all comparisons they made and the parameters of their own portfolios, $r = .299$, $p <.01$. This suggests a relationship between the effort that was exerted in rationalising why one portfolio was of better quality than another, with a student's own performance. This may indicate that students who performed better in the task were better able to describe or distinguish quality in other student's work, potentially indicating a reason for their own increased performance.

The next stage of the analysis examined the nature of the criteria used by judges when making a comparative decision. Out of the 126 students, 91 provided details of the criteria governing the decisions they made. Excluding comparisons where no details of judgement criteria were provided, a total of 1067 criteria were given across 596 comparisons. Due to the considerable overlap in the language within the criteria used, the original 1067 criteria were reduced into 60 unique codes. Finally, these codes were inductively categorised into 15 sub-categories of criteria within six broad categories. Table 3 presents the coded criteria and an analysis of their frequencies. It is important to note that the categories and sub-categories give context to the codes and all three levels are needed to ensure an accurate interpretation. For example, the code 'work quality' could be interpreted to reflect the quality of the entire portfolio however during the coding process it was used specifically in the context of the craft or manufacturing aspect of the project. It was therefore placed in the sub-category 'Quality of work' within the category 'Pure Craft'.

Table 3. Coded judgement criteria with frequencies and z-scores.

Criteria	Frequency	*z*-score
Brief Requirements	8	-1.140
Requirements of brief	8	-.838
Meets brief	8	-.291
Pure Professional Judgement	69	-.730
Professional judgement	69	-.028
Better overall	58	1.197
Impressive project	4	-.410
Personal preference	7	-.321
Portfolio	269	.612
Effectiveness of portfolio	241	2.255*
Story communication	6	-.351
Better portfolio	46	.840
Portfolio communication	16	-.053
Better communication	87	2.061*
Use of media	29	.334
Tagging	17	-.023
Entertaining portfolio	2	-.470
Idea development	1	-.500
More interesting	3	-.440
More sketching	1	-.500
Interesting portfolio	1	-.500
Concept communication	1	-.500
Better language	3	-.440
Portfolio detail	28	.304
Evidence of work/learning	28	-.573
Proof of work	16	-.053
Evidence of learning	5	-.381
Display of knowledge	2	-.470
Display of skills	2	-.470
Connection to lectures	3	-.440
Pure Design	413	1.578
Conceptual design	123	.689
Better concept	44	.781
Good concept	4	-.410
Unique	9	-.262
More creative	1	-.500
Better story	6	-.351

Criteria	Frequency	*z*-score
Risk taken	2	-.470
Use of materials	1	-.500
Interesting theme	4	-.410
Adventurous design	1	-.500
Design detail	20	.066
Better design	28	.304
Interesting story	1	-.500
Better idea	2	-.470
Conveying emotion	163	1.220
Emotion communication	141	3.669**
Personal	16	-.053
More emotion	4	-.410
Deeper emotion	2	-.470
Artefact coherency	43	-.373
Coherent project	36	.542
Theme communication	4	-.410
Project communication	3	-.440
Level of design work	36	-.466
More thought	27	.274
Modelling	6	-.351
Planning	3	-.440
Relatable to assessor	35	-.480
Relatable	35	.513
Artefact appearance by design	13	-.772
Use of colour	13	-.143
Pure Craft	207	.196
Quality of work	203	1.751
Work quality	200	5.425**
Well executed	3	-.440
More skills used	4	-.891
More skills used	4	-.410
Partial Design/Partial Craft	101	-.516
Artefact appearance	43	-.373
WOW	9	-.262
Visually superior	30	.364
More appealing	1	-.500
More 'striking'	3	-.440
Level of work	47	-.320
More work	42	.721

Criteria	Frequency	*z*-score
More effort	4	-.410
Perceived difficulty	1	-.500
Specific element	11	-.798
Specific element	10	-.231
Craft detail	1	-.500

Note. *z*-scores were calculated based on the frequency of a code, sub-category or category relative to the mean and standard deviation of their group. ** *z*-score is significant at the 0.01 level (2-tailed). **z*-score is significant at the 0.05 level (2-tailed).

Based on the frequencies of the various specific criteria used, of those that were mentioned a statistically significant number of times, the most frequently cited rationale governing a judgement was the quality of work (z = 5.425) which was specifically associated with the craft of the artefact. This was followed by the communication of emotion (z = 3.669) which was associated with the design of the artefacts, but it should be noted that this was a requirement of the brief. Finally, a better communicated portfolio (z = 2.061) was frequently cited as the reason one portfolio was deemed better than another. Only one of the sub-categories was cited a statistically significant amount of times, the general efficacy of the portfolio (z = 2.255). Taken together, this suggests that 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.

The final analysis examined the variance in criteria used by participants across their judgements. While 91 students provided commentary on the criteria they used, in order to examine the potential differences in this criteria across judgements, cases where commentary was provided on only one comparison were removed. A total of 79 students provided commentary on more than one judgement and were considered for the following analysis. Rather than considering potential variances in terms of the coded responses, this analysis focused on the sub-categories and categories of codes (Table 3) to control for variance based on language use rather than on judgement rationale. On average, participants provided commentary on the criteria they used on 7.392 judgements (SD = 2.409). In terms of the coded sub-categories, participants used an average of 5.443 (SD = 2.129) unique criteria and in terms of categories they used an average of 3.722 (SD = 1.120) unique criteria.

A detailed analysis of the number of unique decisions, i.e. the criteria used to make a judgement was uniquely different to their other judgements, is presented in Figure 1. With respect to the subcategories of criteria (illustrated in Table 3), for 5.06% of the cohort 33.33% of the judgements were made on unique criteria, while 26.58% of the cohort made 100% of the judgements on uniquely different criteria. In relation to the broader categories of criteria (illustrated in Table 3), for 1.27% of the cohort 22.22% of the judgements were made on unique criteria, while 11.39% of the cohort made 100% of the judgements on uniquely different criteria. These results illustrate quite clearly that there was often a need to make judgements about features of quality or performance on qualitatively different criteria, which reflects the varied and idiosyncratic nature of authentic design and the need for the holistic approach to assessment that is responsive to the work rather than the judge trying to make it fit the rubric.

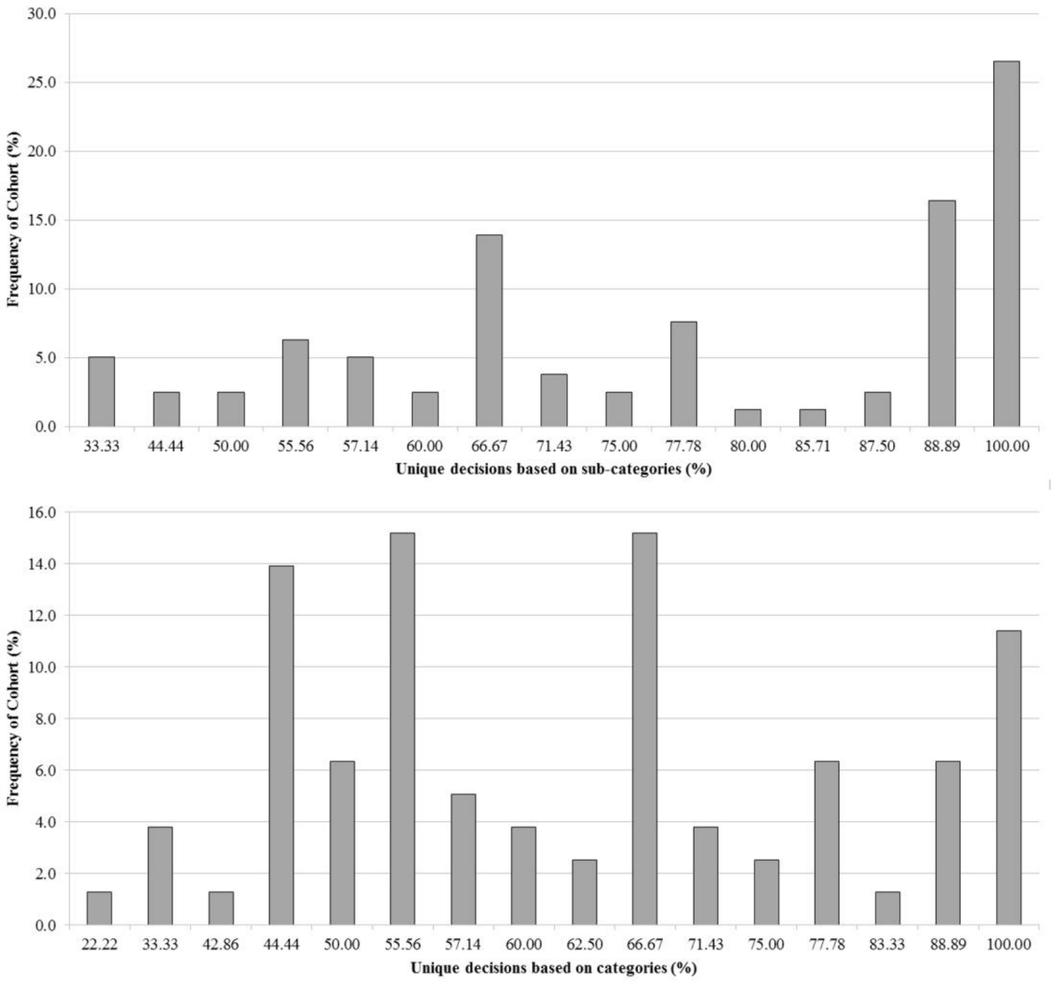

Figure 1. Frequency of unique decisions based on sub-categories (top) and categories (bottom).

Discussion

The results of this study have a number of implications for the assessment of design related outputs in technology education. The main question which needs to be considered, especially in light of the current reform, is whether an assessment approach similar to ACJ would be beneficial. The evidence from this study is clear that the students felt a need to vary the criteria that they used when making an adjudication on quality. Any form of open-ended assignment will result in varied student outputs and thus varied evidence to demonstrate features of quality, and the future work completed by technology students in relation to CBA's and final projects will be no different. The use of ACJ appears a viable option to alleviate the concerns put forward by Sadler (2009) with respect to the use of criterion-referenced assessment.

The next issue is the fairness of such an approach, i.e. should varied criteria be applied? Considering the high level of reliability which is consistently observed through this method (Kimbell 2012; Seery, Canty, and Phelan 2012; Bartholomew et al. 2017; Bartholomew, Strimel, and Jackson 2018; Seery et al. 2018) and which was again observed in this study (α = .974), it is clear that despite the variety of criteria which were applied in judgements, consensus of quality can be achieved. Additionally, by not imposing strict assessment criteria at the beginning of a project, ACJ can liberate students in terms of their design freedom (Seery, Canty, and Phelan 2012) thus making the activity more authentic.

In a similar line of enquiry, if variable criteria are to be used, there is a need to examine what criteria are applied to judgements. In this particular study, the total number of videos that the students made themselves, the number of photographs that they took and included in their portfolios, and the number of online images they included were able to account for 33.3% of the variance in performance. The qualitative results do suggest a level of validity to this result as some decisions were based on criteria such as more sketching (z = -.500), more modelling (z = -.351), more planning (z = -.440), more skills used (z = -.410), more work (z = .721), and more effort (z = -.410). However, these reasons were typically cited below the average number of times and were non-significant. The only significantly frequently cited criteria used were associated with the quality of craft, the conveyance of emotion which was a requirement of the brief, and the efficacy of the portfolio. Each of these are relevant criteria for evaluating the work produced and based on this, it appears that the amount of work may have simply aligned with quality. However further work should be conducted in a more controlled environment, possibly with specifically designed portfolios of varying quality and with varying quantities of work, to gain further insight into the influence of the quantity and quality of work on design output adjudication.

With respect to the mitigation of biases, this study used the tagging system of having, growing and proving ideas as a proxy for how the students completed their own work. While there was no correlation between the use of these tags and performance, there was a significant association suggesting that a student judge was more likely to pick a winning portfolio if it was more different to their own. However, the effect size was very small, and likely to be negligible in a practical implementation, especially when considered in light of the high level of reliability observed.

Overall, ACJ appears like a potentially auspicious mechanism to support the assessment of design related open-ended products of student work in technology subjects. With respect to CBA's, it is possible the students could act as judges which could see improved learning as a result, similar to the work conducted by Seery et al. (2018) with undergraduate students. The misfit statistics would allow for teachers and students to judge the work and determine the level of alignment between both in terms of determinations of quality. This would allow teachers to see whether their students had consensus in what their perception of quality was and if it aligned with their own. This process could have significant pedagogical implications especially considering the finding of this study that students who were better able to express features of quality performed better in the assignment. In terms of a national assessment, cohorts of teachers could be involved anonymously in the assessment of student work with the added benefit of multiple professional judgments and a high level of reliability in the grading process. This would also allow for teachers to gain insight as to their alignment with their peers' perceptions of quality. Finally, the capacity to make adjudications on features of quality based on varied criteria across judgements would liberate teachers and students from imposed assessment criteria, potentially allowing for greater alignment with learning outcomes of innovation and creativity. However, in order to make clearer determinations on this, it would be necessary to trial the use of ACJ with assignments similar to what the new CBA's and final projects will look like with post-primary students and their regular teachers.

References

Bartholomew, S., E. Reeve, R. Veon, W. Goodridge, V. Lee, and L. Nadelson. 2017. "Relationships between Access to Mobile Devices, Student Self-Directed Learning, and Achievement." *Journal of Technology Education* 29 (1): 2–24.

Bartholomew, S., G. Strimel, and A. Jackson. 2018. "A Comparison of Traditional and Adaptive Comparative Judgment Assessment Techniques for Freshmen Engineering Design Projects." *International Journal of Engineering Education* 34 (1): 20–33.

Bartholomew, S., G. Strimel, and E. Yoshikawa. 2018. "Using Adaptive Comparative Judgment for Student Formative Feedback and Learning during a Middle School Design Project." *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-018-9442-7.

Bartholomew, S., and E. Yoshikawa-Ruesch. 2018. "A Systematic Review of Research around Adaptive Comparative Judgement (ACJ) in K-16 Education." In *CTETE - Research Monograph Series*, edited by John Wells, 1:6–28. Virginia, USA: Council on Technology and Engineering Teacher Education.

Bartholomew, S., E. Yoshikawa, E. Hartell, and G. Strimel. 2019. "Identifying Design Values across Countries through Adaptive Comparative Judgment." *International Journal of Technology and Design Education*. https://doi.org/10.1007/s10798-019-09506-8.

Canty, D., N. Seery, and P. Phelan. 2012. "Democratic Consensus on Student Defined Assessment Criteria as a Catalyst for Learning in Technology Teacher Education." In *PATT2012: Technology Education in the 21st Century*, edited by Thomas

Ginner, Jonas Hallström, and Magnus Hultén, 119–25. Stockholm, Sweden: PATT.

Carty, A., and P. Phelan. 2006. “The Nature and Provision of Technology Education in Ireland.” *Journal of Technology Education* 18 (1): 7–26.

Coertjens, L., M. Lesterhuis, S. Verhavert, R. Van Gasse, and S. De Maeyer. 2017. “Judging Texts with Rubrics and Comparative Judgement: Taking into Account Reliability and Time Investment.” *Pedagogische Studien* 94 (4): 283–303.

Jones, I., M. Swan, and A. Pollitt. 2015. “Assessing Mathematical Problem Solving Using Comparative Judgement.” *International Journal of Science and Mathematics Education* 13 (1): 151–77.

Kimbell, R. 2007. “E-Assessment in Project e-Scape.” *Design and Technology Education: An International Journal* 12 (2): 66–76.

Kimbell, R. 2012. “Evolving Project E-Scape for National Assessment.” *International Journal of Technology and Design Education* 22 (2): 135–55.

Kimbell, R., G. Martin, W. Wharfe, T. Wheeler, D. Perry, S. Miller, T. Shepard, P. Hall, and J. Potter. 2005. *E-Scape Portfolio Assessment: Phase 1 Report*. London: Goldsmiths, University of London.

Kimbell, R., T. Wheeler, S. Miller, J. Bain, R. Wright, and K. Stables. 2004. *Assessing Design Innovation: Final Report*. London: Goldsmiths, University of London.

Kimbell, R., T. Wheeler, S. Miller, and A. Pollitt. 2007. *E-Scape Portfolio Assessment: Phase 2 Report*. London: Goldsmiths, University of London.

Kimbell, R., T. Wheeler, K. Stables, T. Shepard, F. Martin, D. Davies, A. Pollitt, and G. Whitehouse. 2009. *E-Scape Portfolio Assessment: Phase 3 Report*. London: Goldsmiths, University of London.

Kurz, A., S. Elliott, J. Wehby, and J. Smithson. 2010. “Alignment of the Intended, Planned, and Enacted Curriculum in General and Special Education and Its Relation to Student Achievement.” *Journal of Special Education* 44 (3): 131–45.

NCCA. n.d. *Leaving Certificate Architectural Technology: Ordinary Level and Higher Level Draft Syllabus*. Dublin, Ireland: The Stationery Office, Department of Education and Science.

NCCA. n.d. *Leaving Certificate Engineering Technology: Ordinary Level and Higher Level Draft Syllabus*. Dublin, Ireland: The Stationery Office, Department of Education and Science.

NCCA. 2007a. *Leaving Certificate Design and Communication Graphics Syllabus*. Dublin, Ireland: The Stationery Office, Department of Education and Science.

NCCA. 2007b. *Leaving Certificate Technology Syllabus*. Dublin, Ireland: The Stationery Office, Department of Education and Science.

NCCA. 2018a. *Junior Cycle Applied Technology*. Dublin, Ireland: Department of Education and Skills.

NCCA. 2018b. *Junior Cycle Engineering*. Dublin, Ireland: Department of Education and Skills.

NCCA. 2018c. *Junior Cycle Wood Technology*. Dublin, Ireland: Department of Education and Skills.

NCCA. 2019. *Junior Cycle Graphics*. Dublin, Ireland: Department of Education and Skills.

Newhouse, C. P. 2014. "Using Digital Representations of Practical Production Work for Summative Assessment." *Assessment in Education: Principles, Policy and Practice* 21 (2): 205–20.

Nicol, D., and D. Macfarlane-Dick. 2006. "Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31 (2): 199–218.

Pollitt, A. 2012a. "Comparative Judgement for Assessment." *International Journal of Technology and Design Education* 22 (2): 157–70.

Pollitt, A. 2012b. "The Method of Adaptive Comparative Judgement." *Assessment in Education: Principles, Policy & Practice* 19 (3): 281–300.

Sadler, D. R. 2009. "Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning." In *Assessment, Learning and Judgement in Higher Education*, edited by Gordon Joughin, 45–63. Netherlands: Springer.

SEC. 2019a. *Materials Technology Wood: Coursework - Design Briefs*. Dublin, Ireland: State Examinations Commission.

SEC. 2019b. *Metalwork: Techniques and Design - Project - Higher Level*. Dublin, Ireland: State Examinations Commission.

SEC. 2019c. *Metalwork: Techniques and Design - Project - Ordinary Level*. Dublin, Ireland: State Examinations Commission.

SEC. 2019d. *Technology: Design Tasks*. Dublin, Ireland: State Examinations Commission.

Seery, N., J. Buckley, T. Delahunty, and D. Canty. 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*. https://doi.org/10.1007/s10798-018-9468-x.

Seery, N., D. Canty, and P. Phelan. 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): 205–26.

Seery, N., R. Kimbell, J. Buckley, and J. Phelan. 2019. "Considering the Relationship between Research and Practice in Technology Education: A Perspective on Future Research Endeavours." *Design and Technology Education: An International Journal* 24 (2): 1–12.

Steedle, J. T., and S. Ferrara. 2016. "Evaluating Comparative Judgment as an Approach to Essay Scoring." *Applied Measurement in Education* 29 (3): 211–23.

Thurstone, L. L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34 (4): 273–86.

Williams, P. J. 2000. "Design: The Only Methodology of Technology?" *Journal of Technology Education* 11 (2): 48–60.