Modelling approaches to combining and comparing independent adaptive comparative judgement ranks

Jeffrey Buckley, Technological University of the Shannon
jbuckley@ait.ie
Niall Seery, Technological University of the Shannon
niall.seery@tus.ie
Richard Kimbell, Technological University of the Shannon
r.kimbell@gold.ac.uk

ABSTRACT

The use of Adaptive Comparative Judgement (ACJ) for educational assessment addresses one need within technology education for the reliable assessment of responses to open-ended activities which are characteristic within the field. The output of an ACJ session is a rank order of the piece of student work with relative "ability scores". However, the use of ACJ has been limited to date in that ranks are not directly comparable. For example, a rank produced from one class group has no reference information against which to compare a rank produced of the work of another class group. In this type of case a solution has been to combine the work of both classes into one ACJ session, but this has limitation when considering scaling up.

A new goal for the use of ACJ involves solving this issue. The ability to compare or merge ranks presents a new capacity for ACJ – to use a rank as a "ruler" against which other ranks can be compared. In practice this would allow for two possibilities. The first is that a single rank could be developed which presents a national standard against which teachers could compare the work of their students to see where they are performing on a national level. The second is that communities of practice could complete ACJ sessions within their own classrooms, and when meeting as a group they could merge and compare relative performance of their own students to support professional development.

In a previous article a proof of concept of this process conducted via simulation was presented (Authors, 2022). In this article we present the results of a project with authentic data – student work completed in response to meaningful activities with teachers acting as ACJ judges – which indicate that the use of ACJ in this way is now possible.

Key Words: Adaptive comparative judgement, assessment, steady state, authentic evidence.

1. INTRODUCTION

The use of adaptive comparative judgement (ACJ) for assessment in technology education was initially proposed and demonstrated through the e-scape project (Kimbell, 2007). The e-scape project aimed to introduce e-portfolios into technology education in a way that would "free learners from the burdens of artificial story-telling and allow them just to get on with their designing" (Kimbell, 2012, p. 136). Within the e-scape project, ACJ then provided an assessment mechanism which would "cancel out" assessors personal standards (Kimbell, 2007, p. 71). ACJ is said to achieve this by having several assessors, or "judges", collectively make binary pairwise comparisons between pieces of student work, or "portfolios, and by collating the decisions of several judges, individual biases are partialled out of the final rank of all work included (cf. Hartell & Buckley, 2021).

Interest in examining the various possible uses and benefits of ACJ for enhancing technology education has grown over the past decade following a special issue on the topic (Williams & Kimbell, 2012), with recent reviews providing an account of the current state of this research endeavour (Bartholomew & Jones, 2022; Buckley et al., 2022). An ongoing agenda is to progress ACJ beyond its current utility, which is limited to individual assessment sessions, and to expand its capacity for national assessment (Seery et al., 2022). This has been a goal ever since it was introduced into technology education (Kimbell, 2012), but recent functional advances are closing this gap (Buckley & Canty, 2022). This paper presents an empirical study which illustrates a new capacity for ACJ, the ability to consolidate and compare unique ranks of student work. Having this capacity would permit the conduction of several, logistically more feasible, small-scale ACJ assessment sessions and then both merging them into a national rank of student work and comparing individual ranks which could represent different geographical jurisdictions. To illustrate a need for this and to contextualise this process, a brief overview of the ACJ method will first be provided.

2. THE METHOD OF ADAPTIVE COMPARATIVE JUDGEMENT

ACJ ultimately involves having several judges collectively produce a rank order for a collection of portfolios which describes the relative best to relative worst pieces of work. Initially, a sample of portfolios and cohort of judges are identified. Typically the process from this point is managed by propriety software (e.g. RM Compare, 2023) where the portfolios are digitised and judges are given individual accounts. The session begins with the portfolios being randomly paired together, with individual pairs being presented to judges. Each judge then makes a binary decision of which portfolio is "better" or "worse". At least in technology education, this decision has typically been made on a holistic construct of capability (e.g. Seery et al., 2019) with judges being shown to make decisions on varying criteria which are personally selected (Buckley et al., 2020), although formal criteria could be used. After a designated number of judgements are made through a Swiss tournament system[1], an adaptive algorithm is initiated to manage the pairing of portfolios. This is

[1] For the Swiss tournament, in the first round portfolios are randomly paired together for comparison. The result will be half of the portfolios having 1 winning result, and the other half having 0 winning results. In

a defining characteristic separating comparative judgement (CJ) from ACJ. The adaptive algorithm pairs portfolios based on which pairings provide the most information in terms of generating the rank. At the end of the session, which can be determined in several ways such as when a certain level of reliability is achieved or after a prescribed number of judgements (e.g. Verhavert et al., 2022), the decisions from each judgement are used to fit a Bradley-Terry-Luce (BTL) model, which generates a rank order of the portfolios included in the session. The BTL model is computed by

$$\alpha_i = \frac{W_i}{\sum_{j \neq i} \frac{w_{ij} + w_{ji}}{\alpha_i + \alpha_j}} \quad 1$$

where *i* and *j* are individual portfolios, W_i is the total number of wins of portfolio *i*, w_{ij} is the number of wins portfolio *i* has against portfolio *j*, w_{ji} is the number of wins portfolio *j* has against portfolio *i*, α_i is the ability score estimate of portfolio *i*, and α_j is the ability score estimate of portfolio *j* (Hunter, 2004). Initially, all ability scores are estimated as 1 and then normalised to maximum likelihood estimates.

Importantly, the rank does not provide any absolute indicators of quality. Performance is denoted in "parameter values" or "ability scores" which are centred around 0 (i.e., a score of 0 represents the theoretical average portfolio, with positive scores being above average and negative scores below average). Taking the top and bottom ranked portfolios as an example, once the rank is generated there is still no determination whether either is "good" or "poor" in terms of absolute performance. The entire rank could represent outstanding work, it could all be very poor quality work, or it could range from anywhere in between. A process beyond the ACJ session is required to map the rank onto, for example, grades which could denote performance. The rank is limited to relative performance indication where the quality of any individual portfolio is only presented as a relative value in comparison to all other portfolios in the rank. This presents a significant limitation in that if two independent ranks are generated, both will have a mean ability score of 0 and within-rank relative ability scores, but these scores are not immediately comparable between ranks. As such, independent ranks cannot be consolidated or compared directly without an additional procedure where they are adjusted onto the same scale. The paper presents a study where three approaches to scaling ranks are explored to alleviate this current limitation.

3. METHOD

Four ACJ sessions were conducted as part of this study. The first three of these were typical ACJ sessions managed through the RM compare (2023) system where the portfolios were paired

the subsequent rounds, portfolios are again randomly paired but now only with those which have the same or a similar number of wins as they have. Thus, during the second round for example, portfolios with 1 win after round one are paired randomly with other portfolios with one win, whereas those with 0 wins are paired with others which also have 0 wins. The outcome of round two being a selection of portfolios with 2 wins, 1 win, and 0 wins. This process then repeats for the designated number of rounds, with no portfolios being paired together more than once.

initially by a Swiss tournament system and then by an adaptive algorithm as previously described. These are herein referred to as Session A, B, and C respectively. The fourth ACJ session (herein Session D) utilised a novel adaption to the RM Compare platform where pairs could be manually determined in advance of the session and assigned to specific judges. In total, there were 13 judges who were all technology education educators at either secondary level or in higher education on technology teacher education programmes, and 35 portfolios which were generated in response to an authentic task by secondary level technology pupils. The portfolios were generated by pupils in two schools, School A and School B. Pupils in School A submitted 17 portfolios with 18 portfolios being submitted from School B. The task was a classroom-based assessment (CBA) which is a new introduction to the Irish lower-secondary school system (Department of Education and Skills, 2015). Twice in students' lower secondary education they must complete a CBA which is then assessed by their teacher for formative purposes only. The results of these are indicated in students' Junior Cycle Profile of Achievement (JCPA), a record of their overall performance at lower secondary level. The task completed by all students was an investigation into "The ergonomics of household objects" and it was a CBA assigned nationally to all students taking the technology subject of Graphics. While responses to a national assessment, no data were collected to provide an indication as to whether the portfolios collected in this study were nationally representative.

All 13 judges participated in each of Session A, B, and C, and the portfolios were assigned to each session as shown in Table 1. Note that portfolios submitted from School A received an anonymous ID in the format portfolio.aX and portfolios submitted from School B received an anonymous ID in the format portfolio.bX.

Table 1.
Portfolio ID's and list of portfolios included in Sessions A, B, and C.

Portfolio No.	Portfolio ID	Session A portfolios	Session B portfolios	Session C portfolios
1	portfolio.a1	●		●
2	portfolio.a2	●		●
3	portfolio.a3	●		●
4	portfolio.a4	●		●
5	portfolio.a5	●		●
6	portfolio.a6	●		●
7	portfolio.a7	●		●
8	portfolio.a8	●		●
9	portfolio.a9	●		●
10	portfolio.a10	●		●
11	portfolio.a11	●		●
12	portfolio.a12	●		●
13	portfolio.a13	●		●
14	portfolio.a14	●		●
15	portfolio.a15	●		●
16	portfolio.a16	●		●
17	portfolio.a17	●		●
18	portfolio.b1		●	●
19	portfolio.b2		●	●
20	portfolio.b3		●	●
21	portfolio.b4		●	●
22	portfolio.b5		●	●
23	portfolio.b6		●	●
24	portfolio.b7		●	●
25	portfolio.b8		●	●
26	portfolio.b9		●	●
27	portfolio.b10		●	●
28	portfolio.b11		●	●
29	portfolio.b12		●	●
30	portfolio.b13		●	●
31	portfolio.b14		●	●
32	portfolio.b15		●	●
33	portfolio.b16		●	●
34	portfolio.b17		●	●
35	portfolio.b18		●	●

The outcome of each session was a rank order of the included pieces of work, and the rank reliability is denoted by the scale separation reliability (SSR) coefficient computed by

$$SSR = \frac{\sigma_{\alpha}^2 - MSE}{\sigma_{\alpha}^2} \quad 2$$

where σ_{α}^2 is the standard deviation of the estimated ability scores squared, and MSE is the mean squared standard error, or the mean of the standard error values after they have been squared. The results for each of these sessions are presented in Figure 1,
Figure 2, and Figure 3 respectively.

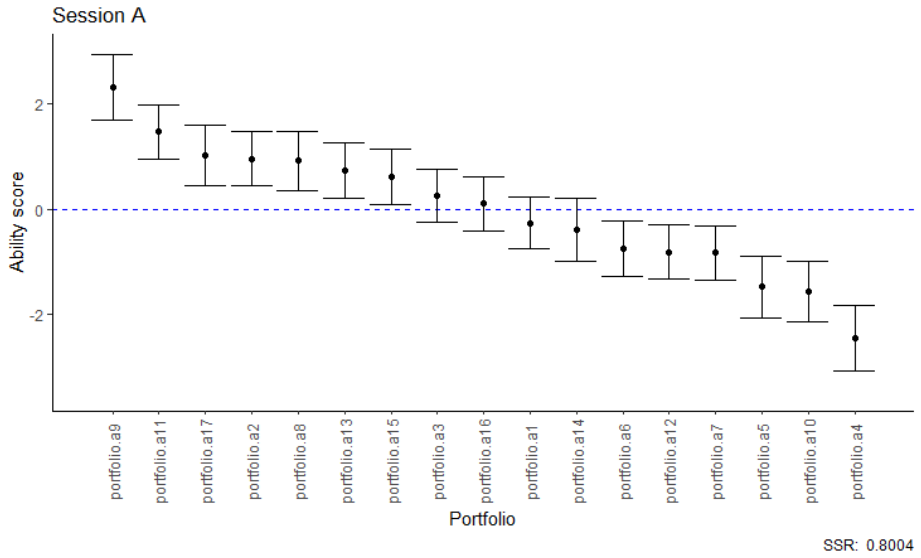

Figure 1.
Session A ACJ rank.

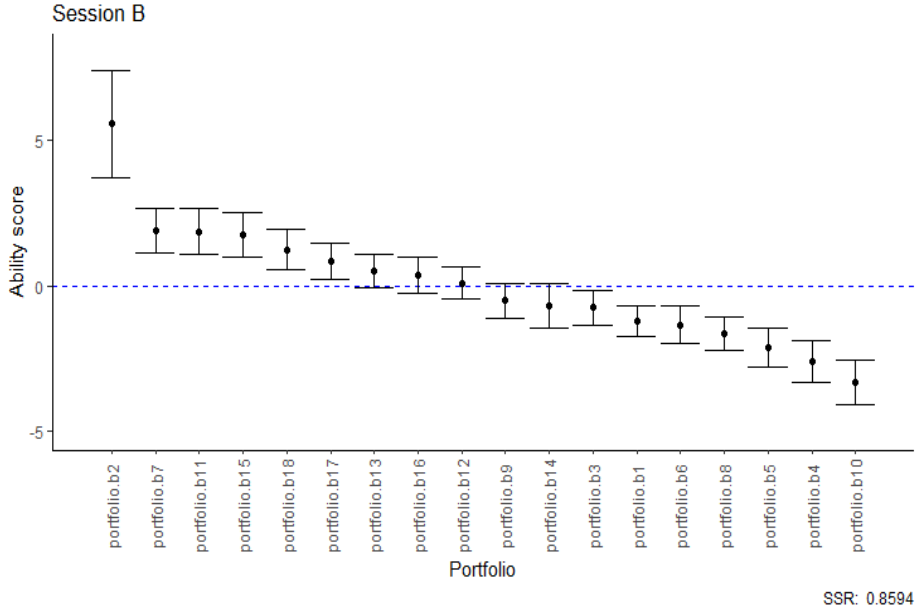

Figure 2.
Session B ACJ rank.

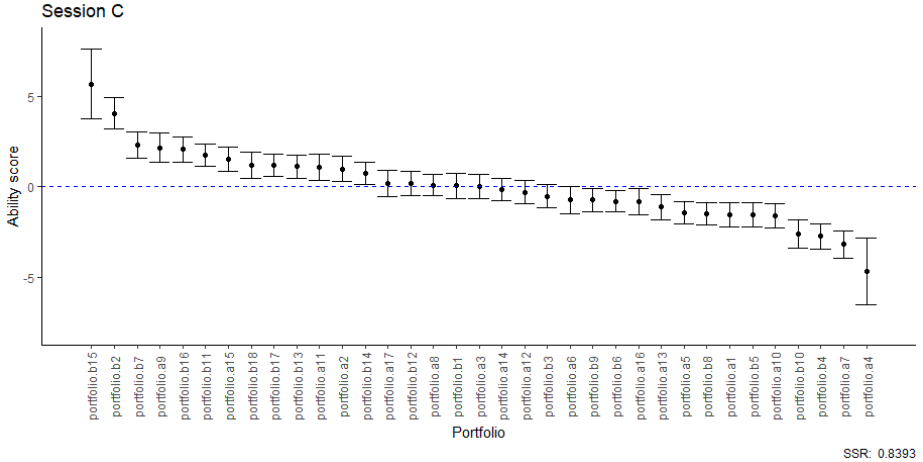

Figure 3. Session C ACJ rank.

Each rank achieved a high level of reliability (SSR > .8). They were also generated with each having a different purpose. The resulting rank from Session C represents a goal state. It contains a single rank containing all portfolios generated through comparative judgements from the judge cohort. It is highly reliable (SSR = .839) and immediately permits the relative performance of School A to be compared with the relative performance of School B. Session's A and B represent separate performance ranks for School A and School B, and in their current states are not immediately comparable nor can they be merged into a single rank. As such, Session D was designed to merge the ranks from Session A and Session B, with the resulting merged rank being compared with that from Session C to see how well the merging process worked. More specifically, three approaches to achieving this were examined and are referred to as Model D1, D2, and D3. In each case, the rank produced from Session B will act as analogous to a proposed "ruler" or steady state concept (Seery et al., 2022). That is, this rank will be fixed, and the achievement of the project aims involve the successful merging of this rank with the rank produced from Session A. As such, the rank from session A will be adjusted through a scaling process to situate it comparably into the rank produced from Session B.

Each of the approaches for Model D1, D2, and D3 followed the same overall process:

1. Select a portfolio(s) from the Session A rank to "judge into" the rank from Session B.

2. Select the portfolios from the Session B rank against which the selected Session A portfolio(s) would be judged.

3. A purposefully selected sample of the original 13 judges would complete the judgements of the portfolios selected in Step 1 and Step 2.

4. Using the judgements from Step 3, produce a rank using the BTL model. In this rank, the "parameter values" or "ability scores" of the portfolios selected from the Session B rank would be fixed to those which were produced through Session B, and thus only the parameter values of the Session A portfolios are recomputed. These recomputed Session A portfolio parameter values would therefore represent the positioning of the selected Session A portfolios within the Session B rank.

5. Using the recomputed parameter values of the selected Session A portfolios, take the judgements made from the original Session A rank, and recompute the entire Session A rank, this time with the recomputed parameter values of Session A portfolios from Step 4 being fixed. This would produce a completely recomputed Session A rank, with parameter values now scaled relatively to those from Session B.

6. Merge the recomputed Session A rank from Step 5 with the original Session B rank.

7. Compute a correlation coefficient from the merged rank from Step 6 with the original Session C rank.

This procedure requires the selection of judges (Step 3) to make the new judgements. These were selected by first getting the mean misfit statistic for each judge based on Session A and Session B. Judges were then ranked based on the absolute difference between their average misfit and 1 (Table 2). Model D1 needed one judge (judge 2 was used). Model D2 needed three judges, (judges 12,7 and 11 were used). Finally Model D3 also needed three judges (judges 3,1, and13 were used). All judgements which were required (described below in Table 3, Table 4, and Table 5) were run through the single Session D which permitted the manual selection of judgements to be made.

Table 2
Judge ranking based on average misfit from Session A and Session B.

Judge	Session A misfit	Session B misfit	Absolute difference between 1 and average misfit
judge.2	0.797566	1.253231	0.025398
judge.12	1.045002	1.093468	0.069235
judge.7	1.076206	1.112101	0.094153
judge.11	0.88119	1.342626	0.111908
judge.3	1.256328	1.006778	0.131552
judge.1	0.679718	0.819701	0.25029
judge.13	0.760399	1.794187	0.277293
judge.5	1.304007	0.138742	0.278625
judge.6	1.603682	0.972127	0.287904
judge.9	0.72778	0.676694	0.297763
judge.10	1.560905	1.419284	0.490095
judge.4	0.704051	0.18031	0.55782
judge.8	0.26236	0.473132	0.632254

4. RESULTS

4.1. Model D1

For this model, the middle (median) portfolio from Session A (Figure 4) was compared to each portfolio in Session B in a random order (Table 3). These decisions represent steps 1-3 from the previously described process. Next, the previously described steps 4 to 6 were completed, which resulted in a recomputed parameter value for portfolio.a16, which was used to scale the resulting Session A portfolio parameter values and then merge these into the original Session B rank. The resulting Pearson (linear/parametric) and Spearman (monotonic non-linear/non-parametric) correlation coefficients between this merged rank and original Session C rank were $r = .82$ [95% CI; .67, .91], $p < .001$ and $\rho = .86$ [95% CI; .74, .93], $p < .001$ respectively, which are very strong, and they resulted from a single linked portfolio and a single judge.

Figure 4.
Model D1 Session A rank portfolio selection.

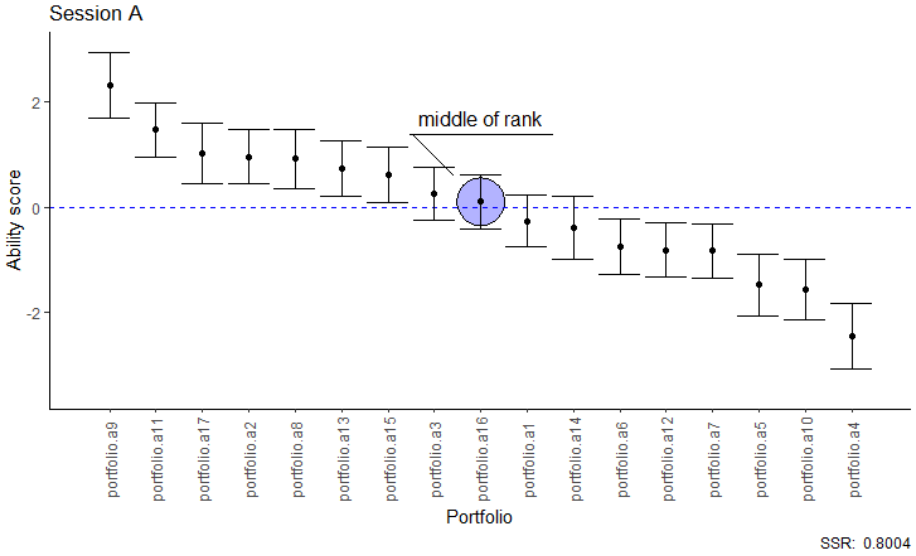

Table 3.
Model D1 comparisons in order.

Session B portfolio	Compared Session A portfolio by judge.2
portfolio.b5	portfolio.a16
portfolio.b18	portfolio.a16

portfolio.b6	portfolio.a16
portfolio.b12	portfolio.a16
portfolio.b13	portfolio.a16
portfolio.b1	portfolio.a16
portfolio.b14	portfolio.a16
portfolio.b15	portfolio.a16
portfolio.b10	portfolio.a16
portfolio.b3	portfolio.a16
portfolio.b17	portfolio.a16
portfolio.b2	portfolio.a16
portfolio.b11	portfolio.a16
portfolio.b16	portfolio.a16
portfolio.b4	portfolio.a16
portfolio.b7	portfolio.a16
portfolio.b9	portfolio.a16
portfolio.b8	portfolio.a16

4.2 Model D2

Following this, we hypothesised that comparisons from a larger number and spread of portfolios in the Session A rank with portfolios from Session B might improve the correlation of the merged rank with that from session C. For Model D2, the top (relative best), middle (median) and bottom (relative worst) portfolios from Session A (Figure 5) were compared to the random sample of portfolios from Session B (Table 4). The process then proceeded identically to that of Model D1. The resulting Pearson and Spearman correlation coefficients were r = .48 [95%; .18, .70], p = .003 and ρ = .47 [95% CI; .15, .70], p = .005 respectively. These are strong correlations however they are markedly weaker than those from Model D1. Given the wider confidence intervals of portfolios at the extreme tails of the rank (portfolio.a9 and portfolio.a4), it is theorized that this result stems from the lower certainty of the positions of the relative best and relative worst portfolios in a rank. By representing the extremes, they do not have portfolios beyond them in the Session A rank which provide relative information of by how much they are the best and worst in the rank. The re-computation of the original Session A portfolios using these portfolios therefore likely introduced additional error due to the higher degree of uncertainty/higher error associated with these portfolios.

Figure 5.
Model D2 Session A rank portfolio selection.

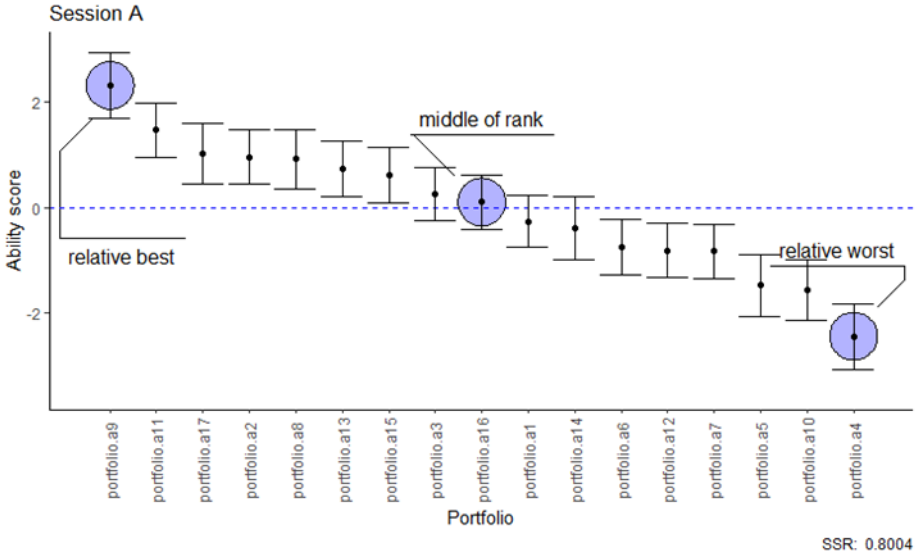

Table 4.
Model D2 comparisons in order.

Session B portfolio	Compared Session A portfolio by judge.12	Compared Session A portfolio by judge.7	Compared Session A portfolio by judge.11
portfolio.b16	portfolio.a9	portfolio.a4	portfolio.a16
portfolio.b5	portfolio.a9	portfolio.a4	portfolio.a16
portfolio.b12	portfolio.a9	portfolio.a4	portfolio.a16
portfolio.b15	portfolio.a9	portfolio.a4	portfolio.a16
portfolio.b9	portfolio.a16	portfolio.a9	portfolio.a4
portfolio.b17	portfolio.a16	portfolio.a9	portfolio.a4
portfolio.b6	portfolio.a16	portfolio.a9	portfolio.a4
portfolio.b4	portfolio.a16	portfolio.a9	portfolio.a4
portfolio.b2	portfolio.a4	portfolio.a16	portfolio.a9
portfolio.b7	portfolio.a4	portfolio.a16	portfolio.a9
portfolio.b18	portfolio.a4	portfolio.a16	portfolio.a9
portfolio.b10	portfolio.a4	portfolio.a16	portfolio.a9

4.3 Model D3

Based on the results of Model D2, to avoid using portfolios from the extremes of the Session A rank, three portfolios down from the top and up from the bottom, and also from the middle were selected from the Session A rank (Figure 6) to merge into the rank from Session B (Table 5). The aim here was to reduce the error introduced by using portfolios at the extreme ends which we believed introduced greater error. With this approach, we build in more information than Model D1, but the relative best and second best and relative worst and second worst mean there are portfolios now on either side of all selected portfolios from Session A which provide additional information regarding relative positioning. The resulting Pearson and Spearman correlation coefficients were $r = .79$ [95% CI; .62, .99], $p < .001$ and $\rho = .80$ [95% CI; .64, .90], $p < .001$. These results are not significantly different from those of model D1 despite the additional information for the rescaling of the Session A rank.

Figure 6.
Model D3 Session A rank portfolio selection.

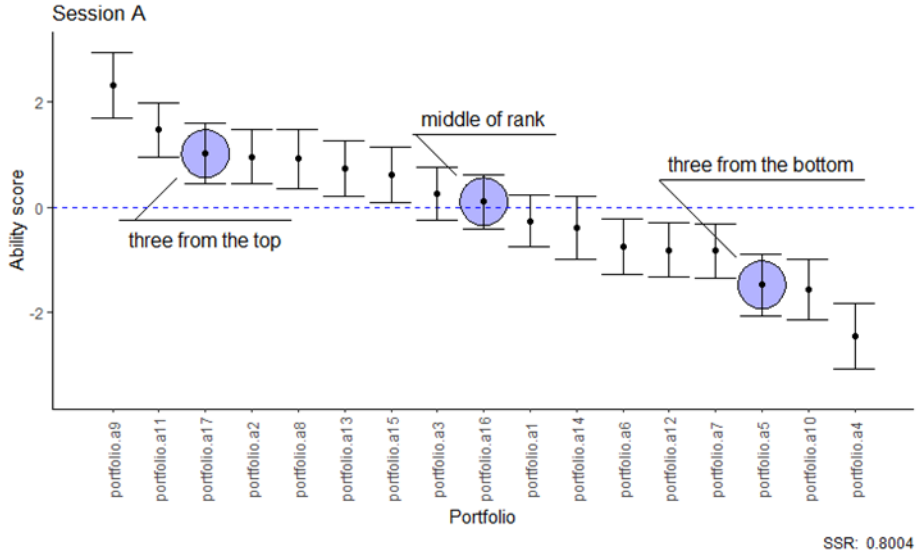

Table 5. Model D3 comparisons in order.

Session B portfolio	Compared Session A portfolio by judge.3	Compared Session A portfolio by judge.1	Compared Session A portfolio by judge.13
portfolio.b16	portfolio.a17	portfolio.a5	portfolio.a16
portfolio.b5	portfolio.a17	portfolio.a5	portfolio.a16
portfolio.b12	portfolio.a17	portfolio.a5	portfolio.a16
portfolio.b15	portfolio.a17	portfolio.a5	portfolio.a16
portfolio.b9	portfolio.a16	portfolio.a17	portfolio.a5
portfolio.b17	portfolio.a16	portfolio.a17	portfolio.a5
portfolio.b6	portfolio.a16	portfolio.a17	portfolio.a5
portfolio.b4	portfolio.a16	portfolio.a17	portfolio.a5
portfolio.b2	portfolio.a5	portfolio.a16	portfolio.a17
portfolio.b7	portfolio.a5	portfolio.a16	portfolio.a17
portfolio.b18	portfolio.a5	portfolio.a16	portfolio.a17
portfolio.b10	portfolio.a5	portfolio.a16	portfolio.a17

5. DISCUSSION

The results of this work are promising in that through this project previously unique ranks which were internally relative were successfully merged. This opens up considerably more functionality for ACJ both for large scale assessment and research purposes as comparative work is now more possible. By supporting comparisons between ACJ ranks of student work, this functionality could also benefit professional development for teachers as they could see and discuss how students work is comparable on larger scales than before. The merging of ranks required additional comparisons to be made, which were managed through a more controllable version of ACJ developed by RM Compare, and working with the BTL model outside of existing ACJ software systems. It should be noted that this project could also have been designed such that rather than scaling the Session A rank to fit into the Session B rank which remained fixed, both the original Session A and Session B ranks could have been merged by fitting the BTL model to the original judgements of both ranks with the new judgements in a single step. This would have meant that both original ranks would have been adjusted. This may be a valuable approach to take in the future, however for this project by fixing one rank in place we demonstrate the functionality not only to merge and thus compare disparate ranks, but also to track relative changes over time by keeping historic ranks fixed for comparability purposes. While the study was exploratory, it appears that working with portfolios at the extremes is not an optimal approach. As such, while future confirmatory studies are important, as this project scales to larger sample sizes where more models can be explored, focusing on portfolios with lower standard error values seems like a strategic choice.

6. ACKNOWLEDGMENTS

The authors wish to express their utmost gratitude to all of the teachers who engaged with us during this project either by acting as judges or by providing anonymised portfolios to be included in the ACJ sessions.

7. REFERENCES

Bartholomew, S. R., & Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. International Journal of Technology and Design Education, 32(2), 1159–1190. https://doi.org/10.1007/s10798-020-09642-6

Buckley, J., & Canty, D. (2022). Assessing performance: Addressing the technical challenge of comparing novel portfolios to the 'ACJ-Steady State'. PATT39: PATT on the Edge - Technology, Innovation and Education, 523–537.

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

Buckley, J., Seery, N., & Kimbell, R. (2022). A review of the valid methodological use of adaptive comparative judgment in technology education research. Frontiers in Education, 7(787926), 1–6. https://doi.org/10.3389/feduc.2022.787926

Department of Education and Skills. (2015). Framework for Junior Cycle 2015. Department of Education and Skills.

Hartell, E., & Buckley, J. (2021). Comparative judgement: An overview. In A. Marcus Quinn & T. Hourigan (Eds.), Handbook for Online Learning Contexts: Digital, Mobile and Open (pp. 289–307). Springer International Publishing. https://doi.org/10.1007/978-3-030-67349-9_20

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. The Annals of Statistics, 32(1), 384–406. https://doi.org/10.1214/aos/1079120141

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), 135–155. https://doi.org/10.1007/s10798-011-9190-4

RM Compare. (2023). RM Compare. RM Compare. https://compare.rm.com/

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). 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(4), 701–715. https://doi.org/10.1007/s10798-018-9468-x

Seery, N., Kimbell, R., & Buckley, J. (2022). Using Teachers' Judgments of Quality to Establish Performance Standards in Technology Education Across Schools, Communities, and Nations. Frontiers in Education, 7. https://www.frontiersin.org/article/10.3389/feduc.2022.806894

Verhavert, S., Furlong, A., & Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. Frontiers in Education, 6. https://www.frontiersin.org/article/10.3389/feduc.2021.785919

Williams, P. J., & Kimbell, R. (Eds.). (2012). Special issue on e-scape [Special issue]. International Journal of Technology and Design Education, 22(2).