International Journal of Technology and Design Education (2019) 29:701–715
https://doi.org/10.1007/s10798-018-9468-x

Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels

Niall Seery[1,2] · **Jeffrey Buckley**[1] · **Thomas Delahunty**[3] · **Donal Canty**[4]

Accepted: 3 September 2018 / Published online: 3 October 2018

Abstract
Educational assessment has profound effects on the nature and depth of learning that students engage in. Typically there are two core types discussed within the pertinent literature; criterion and norm referenced assessment. However another form, ipsative assessment, refers to the comparison between current and previous performance within a course of learning. This paper gives an overview of an ipsative approach to assessment that serves to facilitate an opportunity for students to develop personal constructs of capability and to provide a capacity to track competence based gains both normatively and ipsatively. The study cohort (n = 128) consisted of undergraduate students in a Design and Communication Graphics module of an Initial Technology Teacher Education programme. Four consecutive design assignments were designed to elicit core graphical skills and knowledge. An adaptive comparative judgment method was employed to rank responses to each assignment which were subsequently analysed from an ipsative perspective. The paper highlights the potential of this approach in developing students' epistemological understanding of graphical and technological education. Significantly, this approach demonstrates the capacity of ACJ to track performance over time and explores this relative to student ability levels in the context of conceptual design.

Keywords Ipsative assessment · Adaptive comparative judgment · Constructs of capability · Design education · Technology education

✉ Niall Seery
niall@kth.se

[1] KTH Royal Institute of Technology, Stockholm, Sweden

[2] Athlone Institute of Technology, Athlone, Westmeath, Ireland

[3] University of Nebraska-Lincoln, Lincoln, NE, USA

[4] University of Limerick, Limerick, Ireland

Introduction

Internationally, the measurement of academic learning outcomes has become increasingly important in recent years (Coates 2014) resulting in an urgent need to develop instruments that enable the fair and valid assessment of student competencies (Zlatkin-Troitschanskaia et al. 2015, 2016). Competencies are broadly defined as an amalgam of cognitive, affective, motivational, volitional, and social dispositions underpinning performance (Shavelson 2013) and are assumed to be multidimensional and discipline specific (Zlatkin-Troitschanskaia et al. 2016). While it is of paramount importance to ensure that assessment and evaluation mechanisms are both valid and reliable, these two traits alone are not enough as the structure and nature of assessment can have profound effects on learners (Vaessen et al. 2017). The effects of assessment, from both pedagogical and psychological perspectives, are well documented with notable attributes being affected such as the learning process (Hattie and Timperley 2007), assessment related anxiety (Huxham et al. 2012), self-esteem (Betts et al. 2009), and approaches to learning (Reeves 2006). One commonly used method to alleviate some of these negative effects created through assessment processes is the adoption of a continuous assessment model (Holmes 2014). Additionally, assessment can be incentivised through the provision of feedback which can positively affect learning gains (Black and Wiliam 1998) and, if synthesised appropriately, can support student integration into the assessment process further facilitating positive educational outcomes (Nicol and Macfarlane-Dick 2006). The work of Nicol (2007) supports the view of actively integrating students into the assessment process through engagement with feedback. Nicol (2007) outlines the importance of active participation that supports engaging with diverse views on evidence of learning, and the benefit of evaluating and rehearsing internalised knowledge and understandings when reviewing and preparing feedback. The responsibility is placed on the learner to comprehend the qualities and standards associated with expected performance. This enables a self-regulatory capacity that is critical for learning (Zimmerman 1990). Together with the benefits of engaging with feedback and formative assessment, students must become inducted into the assessment practice (Sadler 2009).

Integrating students into formative assessment

Educational assessment can serve multiple purposes and as such the discourse pertaining to defining the functions of assessment in response to a targeted agenda becomes critical. Functions of assessment mechanisms include the provision of summative evaluations, formative advice to guide learners, and diagnostic information for educators concerning student development.

While teachers may not always share their assessment agenda with their students while enacting a particular practice, ultimately in time it should become clear. This is particularly true in cases where the function of assessment is feedback as despite teaching staff claiming to give good feedback, Hughes (2011) highlights that students often disagree with such claims. Considering the potentially profound effect feedback can have on learning (Black and Wiliam 1998), there is substantial evidence illustrating that it is often poorly delivered resulting in little practical impact (Carless et al. 2011; Price et al. 2010; Scoles et al. 2013). Reasons cited pertaining to this lack of impact include vague feedback (Weaver 2006), not understanding (Lea and Street 1998) or misunderstanding (Carless 2006) feedback, or receiving feedback at a time perceived to be too late such that it is no longer

important (Price et al. 2010). These issues could be reduced or alleviated by sharing the formative agenda of the assessment practice more clearly with students such that they are integrated into the feedback process. Involving the learner in creating feedback as much as they engage in constructing knowledge can satisfy the alignment of this action. Nicol and Macfarlane-Dick's (2006) seven principles of effective feedback support this participatory view of learner involvement.

When assessment is designed with a formative agenda, feedback can also be provided indirectly to learners who are integral to the assessment process. Where students have to assess work themselves, particularly where there is a relatively high degree of subjectivity such as with open ended design tasks, students can develop constructs of quality and capability (Canty et al. 2012; Kimbell 2009; Kimbell et al. 1991). Sadler (2009) argues that the development of a conceptualisation of quality is critical to consistency in high achievement and, if designed accordingly, assessment mechanisms can provide a vehicle for students to develop such constructs. In particular, the adoption of a peer analysis model can facilitate the formulation of constructs of capability through exposure to a variety of examples of work of various levels of quality (Sadler 1989). In contrast, exposing learners to single model solutions or exemplars may make it difficult to develop conceptions of high quality work due to a lack of comparatives limiting goal awareness (Reinholz 2016). Providing an opportunity to critique various solutions to a problem or task allows for strengths and flaws in work to be identified helping to develop a deeper conceptual understanding of quality (Swan 2006). There are multiple mechanisms in existence which can facilitate peer feedback of this nature. Examples include systems with assessment rubrics such as the Calibrated Peer Review (CPR) system (Robinson 2001), and systems where decisions on work are based on holistic judgements such as with Adaptive Comparative Judgement (ACJ) (Pollitt 2012b). In an attempt to integrate the goal of learners developing constructs of capability this paper advocates the use of ACJ. This position stems largely from the work of Sadler (2009), in particular his criticisms of assessment rubrics in that the sum of individual scores may not be representative of the holistic and intuitive opinion of the professional educator, and that learners may produce extraordinary work with unforeseen characteristics not accounted for in pre-designed rubrics whereby the characteristics which identify the work as extraordinary are not appreciated in the assessment mechanism.

Adaptive comparative judgement (ACJ)

Boud and Falchikov (2007) identify a fundamental problem with the dominant discourse in assessment being the positioning of the learners as passive subjects to be measured or classified by the assessment acts of others. The adoption of ACJ can alleviate this concern as learners can be integrated into the assessment process in the form of 'judges' who make holistic decisions on the work of their peers. As Pollitt (2012b) provides a detailed and comprehensive account of the underpinning mathematics of ACJ, this space will be reserved instead for the provision of a more general overview of ACJ with a specific emphasis on its integration within educational settings.

Holistic assessment, as made possible by the ACJ pairs engine (Kimbell 2012; Pollitt 2012a, b) is the judgement of value of 'the whole' rather than the sum of a set of individual components of a task. The ACJ system is premised on Thurstone's (1927) Law of Comparative Judgement and is strengthened by the consistent finding of high reliability scores. ACJ reliability is denoted in a coefficient equivalent to Cronbach's alpha (α) and has been observed in multiple studies to range from $\alpha = .93–.97$ (Baker et al. 2008; Bartholomew

et al. 2017; Bartholomew et al. 2018a, b; Newhouse 2014; Pollitt 2012b; Seery et al. 2012). The system has three elements; (1) a set of portfolios of work created by learners in a response to a task or problem, (2) a community of judges who assess the portfolios through holistic comparisons, and (3) a 'pairs engine' which is a software solution that dynamically selects pairs of portfolios and presents them to judges for adjudication on the quality of the work. The judging process then ultimately results in a rank order of the included portfolios with parameter values denoting the relative distance between them.

A number of critical considerations must be made prior to the ACJ process. The primary components of the assessment process are the pieces of work created by the learners as evidence of their learning which are described in this paper as portfolios for consistency (cf. Kimbell 2012). Portfolios can consist of any digitally formatted pieces of work such as image, text, video or audio files. In relation to the community of judges, the nature or demographic of the involved judges must be considered. From a pragmatic perspective anyone can act as a judge, however their relationship with the assessment task combined with the intended purpose of the rank will ultimately affect the validity and utility of the judgements. Finally, the agenda of the rank must be considered. The rank, while relative in its initial state, can be transformed into academic grades in multiple ways, can be used as both a summative or formative tool, or as is the case in this paper, it can be used to illustrate the ipsative development of learners.

In practice, judges are presented with two portfolios on which they have to make a binary decision on quality. Typically, these decisions are based on the holistic opinion of the judges however there is potential for explicit criteria to be provided in this process. As judges make decisions on pieces of evidence, the process begins with a rough sorting mechanism known as a 'Swiss tournament' system (Pollitt 2012b) to establish broad categories of quality and evolves into an adaptive system whereby pairs for comparison are selected based on criteria such as a similarity in quality or where there is contention on the position of a portfolio. This adaptive nature of the underpinning algorithm is the differentiating characteristic between ACJ and comparative judgement (CJ).

The binary decision required when making pairwise comparison allows judgements to be made on an unarticulated recognition of qualities. As ACJ is typically used in the assessment of open ended design tasks (Bartholomew et al. 2018; Seery et al. 2012), the variance in solutions means judges often employ a tacit understanding of quality when appraising the work. As such, it can be difficult to articulate explicitly what the differentiating characteristics are between portfolios (cf. Newman 2017) and employed criteria may change between judgments. The unarticulated recognition of qualities refers to this process of judging where it is potentially beyond the capacity of judge to explicitly qualify their decision, despite potentially having a tacit understanding of their rationale.

The flexibility in the ACJ approach requires that the judge call on construct of capability to make a value judgement rather than being bound by fixed and predetermined criteria (Hager and Butler 1996). For a learner as a novice or 'quasi-expert' (Kaufman et al. 2013), the judging process forces them to build a conception of what it means to be capable and, as discussed, this can be achieved, at least in part, by exposure to exemplars and the breadth of interpretations and submissions. These qualities must be internally processed by the judge, based on personally set but externally influenced criteria and standards. In addition, subsequent to making a judgement, the judge is prompted to provide feedback to the learner regarding their portfolio and also to provide commentary on the criteria underpinning their decision. This serves two purposes. First it supports the judge developing a personal construct of capability as they externalise their thoughts pertaining to the evidence of learning they were presented with. Second, it serves as a feedback mechanism for the

learners who created the portfolios as an array of commentary on their work is generated which they can review and reflect on. As discussed in relation to open ended design problems, such articulation may be difficult or even impossible to provide accurately so while this commentary can provide insight, it is important that it doesn't bias a judge's decision.

Finally, Bartholomew and Yoshikawa-Ruesch (2018, p. 24) through a systematic review found that "the majority of ACJ research has emphasized summative assessment techniques and has shown positive results. Preliminary efforts into utilizing ACJ as a formative tool for assessment and feedback [have] shown promise and further efforts in both summative and formative applications of ACJ will strengthen arguments related to ACJ's potential for educational transformation". It is therefore clear that efforts aspiring to examine the potential benefits of ACJ with a function as a formative tool are warranted. To date, it has not been applied to observe an individual's personal gain over time. As such, there is a need to explore its implementation within education from a perspective of ipsative enquiry.

Ipsative development

The ipsative approach to assessment was initially conceived by Cattell (1944) where he described it as "a convenient [term] for designing scale units relative to other measurements on the person himself" (p. 294). Ipsative assessment has subsequently been adopted in many studies concerning human intelligence (McDermott et al. 1990) and has since transcended into educational settings as an approach to measure individual learner development. This concept of reference to the self makes the learners progress explicit. This form of feedback qualifies the development over time in response to a target or goal. In practice, the learners' performance in an initial task is then compared to a subsequent task in a sequential process of feedback and action. Developmental tasks can be intermediate and also examined through an ipsative perspective. The benefits of an ipsative approach are multifaceted. From a learners perspective there is a potential motivational value in clearly seeing progress and development relative to a previous achievement or understanding (Hughes 2011). From a teacher's perspective, it illustrates performance over time and allows for a diagnostic scheme to structure learning with defined expectations and/or pathways for individual learners.

By its nature, ipsative assessment supports a cumulative understanding of performance developed over sequential tasks and activities. Tasks that support ipsative assessment tend to be designed to reflect governing principles and a macro view of capability, therefore allowing the learner to compare performances while building associated attitudes, skills and knowledge. Ultimately the approach supports a formative agenda that is largely self-regulated (Hughes et al. 2014). Ipsative assessment also facilitates an effective synthesis of feedback and feed-forward processes (Hughes 2011). It can increase the likelihood of the learner acting on feedback as it is personally relevant and benchmarked with previous performance. However, the capacity of the individual to unpack performance deficits may be outside the capacity of a novice learner especially if there is no reference, either absolute or normative, to expected progress. Therefore, the ipsative agenda must be critically managed. As outlined by Hughes (2011) one of the challenges with ipsative assessment is in operationalizing it. The requirement to design and implement a sequence of learning tasks or activities can be difficult within a modularised and semester based system, usually constrained by approximately 13 weeks of teaching. Additionally, providing feedback that is timely, relevant and bespoke to each learner on all tasks is impractical and further inhibits its implementation.

Method

Approach

In an attempt to mediate many of the difficulties and concerns associated with educational assessment, and particularly ipsative assessment, this study presents a synthesis of ACJ with an ipsative style approach to assessment. Through the use of ACJ learners can become integrated into the assessment process, become self-regulating and be provided with an opportunity to develop constructs of capability through holistic judgement and peer analysis. ACJ can also provide timely feedback through exposure to the work of peers. Finally, as the students themselves can create portfolios of work in response to a design task, the feedback received through the process of making judgments on peers work is both bespoke and relevant to them as individual learners.

As the ACJ process results in a rank order of individual pieces of work, providing individual feedback within a purely ipsative paradigm is immediately possible in practice. However, this study was underpinned by a research question asking how learners of different ability levels progress ipsatively through a series of open ended conceptual design tasks when receiving feedback solely through the process of making pairwise comparisons on the work of peers through the ACJ system. As such, rather than adopting a purely ipsative model which describes individual development, this study examines groups of students by categorising them in terms of quartiles in an initial assignment and subsequently tracking each group across three subsequent assignments. The study employed four assignments over a 12 week semester, all equispaced to observe students' gain over time. Students responded to the design assignments and then, using the ACJ method, made pairwise comparisons on the work both producing a rank order of performance and receiving feedback by virtue of exposure to portfolios of varying standards.

Participants

The study cohort included 3rd Year Initial Technology Teacher Education (ITTE) undergraduate students (N = 136). Students participated in this study as part of a core Design and Communication Graphics (DCG) module which focused on the development of both problem solving skills associated with descriptive and analytical geometry, and graphical communication skills mediated through engagement with conceptual and thematic design assignments. Natural attrition accounted for eight students and therefore the final study cohort consisted of 128 students, of which 123 were male and five were female.

Design

Considering the agenda of measuring competence based gains, it is important to consider the discipline specific nature of such competencies (Zlatkin-Troitschanskaia et al. 2016). The discipline of design education is examined within this study and the competencies inherent within this field, in particular those associated with conceptual design, are speculative in nature (Seery 2017) resulting in diverse and largely unbounded solutions to design challenges. Building on the work of Sadler (2009), this study was cognisant that the prescribed task design was meaningful and pre-planned to support students in evaluating the

quality of evidence created, building appraisal skills, relating directly to learning experiences, and learning how to make judgments about the quality of emerging and finished works holistically rather than based on predefined criteria.

DCG, as a Technology subject, aspires to develop technological knowledge in students (Buckley et al. 2018) which is context specific and largely concerned with application (de Vries 2016). Students were learning explicit graphical knowledge relating to plane and descriptive geometry. Specifically, they were studying four topics including the intersection of geometry, surface developments, lines and planes in space, and conic sections. Four conceptual design assignments were designed and administered which required students to learn and apply knowledge of these topics. The first design assignment required students to apply knowledge of intersecting geometry to synthesise two unrelated objects to create a new and innovative design. The second assignment required the conceptual design of a sports complex to house a sport of the students' choice and to demonstrate their knowledge of surface developments. The third assignment required students to apply their knowledge of lines and planes in space in the proposal of a new conceptual packaging solution for a mobile phone. The final assignment required students to modify a car of their choice in relation to the geometric properties of conic sections. The assignments were underpinned by broad criteria including (1) quality and coherency in communication, (2) innovation, (3) knowledge of stages of design, (4) knowledge of functions of design, and (5) an understanding of geometric principles. A separate study conducted by Seery and Buckley (2016) determined that performance in a similar conceptual graphical design task is indicative of learning and competency with respect to these criteria.

Implementation

Initially, students engaged with the module for 2 weeks prior to receiving their first design brief. During this time, taught elements of the module were delivered to explain module expectations relative to the design assignments and to provide students with information relative to graphical problem solving. Subsequent to this initial instruction period, the four design assignments were administered consecutively with students being allocated 2 weeks for each one. The assignments served as opportunities to learn and apply newly acquired knowledge to resolve a design problem as previously described. An ACJ session was completed at the end of each assignment prior to the beginning of the next. Critically, the students acted as judges during these sessions and were therefore immediately exposed to the work of their peers and as such saw a breadth of interpretation and solutions. On average, students made 8.695 (SD = .590) judgments during each of the four ACJ assessment activities.

To generate grade awards from the ACJ defined rank order, Pollitt (in Kimbell et al. 2009) noted that a suitable linear transformation that preserves the interval scale of the portfolio parameter values generated by the rank order is valid. This requires the setting of two appropriate performance thresholds to guide the grade distribution, i.e. the highest and lowest ranking portfolios must be graded and the individual parameter values for each portfolio are then transformed into grades based on these. In this study, for all assignments, the rank to grade conversion was validated by the module leaders as an independent assignment to ensure validity of performance grades. The academic staff involved in the module (N = 2) graded the assignments independently and then, through a collaborative process, came to a consensus ultimately determining the definition of a standard of performance.

Once the five highest and lowest ranking portfolios were assessed by the academic staff, grades were transposed based on the parameter values to the remaining portfolios.

Results

Prior to examining the performance of each quartile over time in response to the underpinning research question, the initial stage of the analysis involved examining the descriptive statistics associated with each of the assignments (Table 1). An overview of these statistics illustrates that as a cohort, performance improved substantially between the first and last design assignment. As a group, there is also a period between assignments 2 and 3 where performance remains static. The reliability coefficients are higher than those observed in previous ACJ studies (Baker et al. 2008; Pollitt 2012b; Seery et al. 2012) demonstrating a shared understanding of what was of value when appraising the evidence from each of the assignments.

To facilitate the data being analysed from an ipsative perspective relative to performance groups, the cohort was divided into quartiles based on performance on the first assignment (Q1 = 42.40%, Q2 = 55.52%, Q3 = 66.16%, Q4 = 100.00%). Each quartile contained 32 students. The performance in each assignment for each quartile is illustrated in Fig. 1.

A multivariate analysis of variance (MANOVA) was conducted to explore the relationships between performance in each of the four assignments across each quartile. The results indicated a statistically significant difference in assignment scores between quartiles, $F\ (12,\ 320.427) = 21.362$, $p < .000$; Wilk's $\cdots = .211$, partial $\hbar^2 = .405$. To

Table 1 Descriptive statistics

Assignment	n	Mean (%)	SD (%)	Skewness	Kurtosis	α
1	128	54.925	20.131	.095	−.218	.974
2	128	61.797	15.182	−.238	−.290	.973
3	128	60.643	17.062	−.146	−.141	.965
4	128	76.564	10.720	.051	−.290	.971

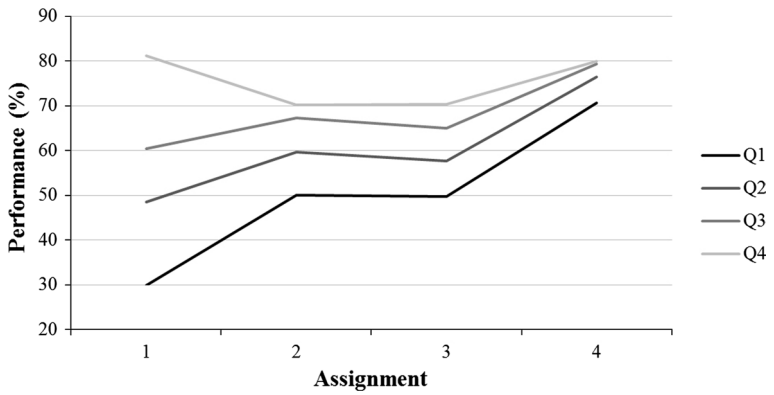

Fig. 1 Performance across each assignment for each quartile

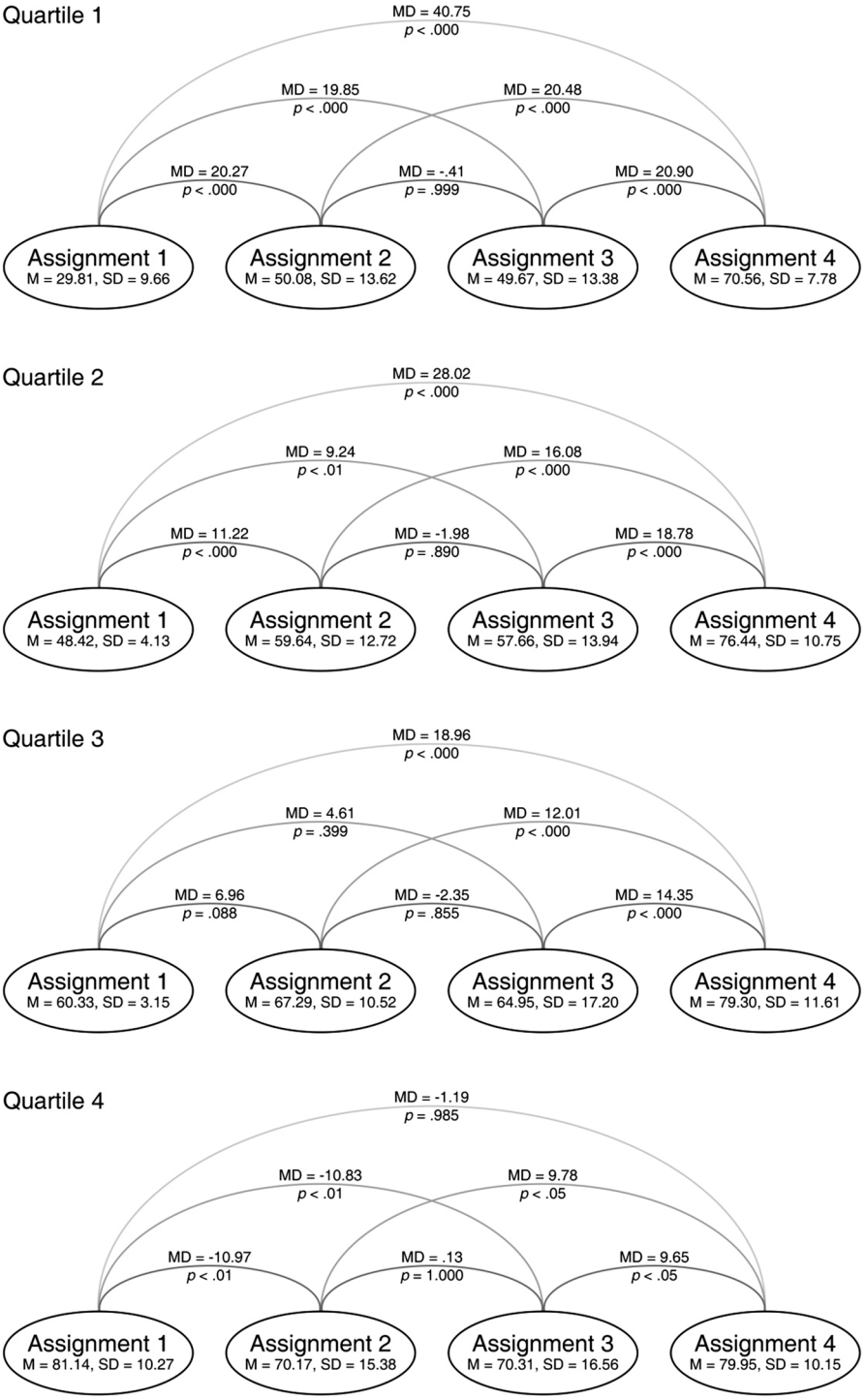

Fig. 2 Tukey's HSD post hoc test results

determine how each quartile performed across assignments, a Tukey's Post Hoc test was used. Figure 2 illustrates the consistent statistically significant difference that existed between Assignment 1 and Assignment 2 (with the exception of quartile 3) and between Assignment 3 and Assignment 4, with no difference in performance between Assignment 2 and Assignment 3 recorded for all baseline quartiles. Although the design of the assignments was similar in principle, and the associated transposition of the ranks to grades was constant, task design and relevant knowledge could account for the differences between assignments.

The tracked improvement for the quartile 1 group shows a gain of 40.75% between Assignment 1 and Assignment 4, while quartile 4 decreased by 1% over the same assignments. While quartile 4 saw a decrease in mean score, it is posited that this may still represent development as relative to Assignment 1, Assignment 4 would have had inherently higher expectations.

From an ipsative perspective how much variance occurred by definition of quartiles as the module progressed is of interest, i.e. did students move into other quartiles? To examine this, the quartiles determine from Assignment 1 were compared with quartiles derived from the results of Assignment 4. These results are presented in Table 2. They indicate that almost half of the students who initially begin in quartile 1 were still in the lowest quartile at the end of the module however one student did ultimately end up in the top quartile. Students who began in the second quartile appear to have evenly dispersed among all quartiles by the end of the semester. The dispersal of students in quartiles three and four appear almost identical. However, slightly more than average ended up in the top quartile with slightly less ending up in the lowest one.

While the previous analyses aimed to address the research question by exploring performance over time relative to baseline ability groups, it was of interest to examine the average amount of time taken to complete judgments in each of the four assignments. This was both to consider the pragmatic application of ACJ and to explore potential differences in judge behaviour across assignments. There was a statistically significant difference between assignments as determined by a one-way ANOVA ($F(3510) = 18.266$, $p < .000$). A Tukey's post hoc test revealed that the average time taken to complete judgments was statistically significantly lower in Assignment 1 ($M = 154.240$, $SD = 117.192$) than in Assignment 3 ($M = 237.377$, $SD = 130.484$, $p < .000$) and Assignment 4 ($M = 259.770$, $SD = 137.395$, $p < .000$). Additionally, the average time taken to complete judgments was statistically significantly lower in Assignment 2 ($M = 191.361$, $SD = 111.584$) than in Assignment 3 ($p = .015$) and Assignment 4 ($p < .000$). There was no statistically significant difference between the time taken to complete judgments between Assignment 1 and Assignment 2 ($p = .075$) or between Assignment 3 and Assignment 4 ($p = .482$).

Table 2 Student movement between quartiles

Initial quartile	Final quartile			
	Q1	Q2	Q3	Q4
Q1	15	8	8	1
Q2	7	9	9	7
Q3	5	8	7	12
Q4	5	7	8	12

Discussion

Reflecting much of the literature on ACJ (Kimbell 2012; Seery and Canty 2017; Seery et al. 2012) the results of this study demonstrated the capacity of ACJ to establish what is of value as determined by a community. In this study, despite the lack of any explicit criteria, the consistently high level of reliability across the four assignments is indicative of the students' agreement on qualities and standards. The level of agreement is evidence that all students were able to recognise what is of value when judging the quality of work irrespective of their own ability level. This indicates that for the lower performing students, their comparatively lower performance was independent of their capacity to recognise higher quality work.

Synthesising the creation of evidence with its appraisal engaged students in a double looped system (Argyris 1976) of reflection in action. The increase in performance across assignments indicates that students were receiving feedback to support them in improving their work. As this only came from the ACJ judging process, this suggests that students were critiquing their own work relative to the breadth of work presented by their peers and they were also engaged in a critique of the purpose of the design assignments with respect to core competency development. In essence, students were developing, responding to, and applying criteria. As students were engaged with the recurring evidence as presented by their peers, it is posited that they developed a better capacity to discriminate between work and as a result formulated standards. This hypothesis is supported by the statistically significant increase in performance between Assignment 1 and Assignment 4 for the entire cohort. This development is often never visible and overall standards are only apparent to a grader at the summative stage of the process.

Two significant elements of this ACJ approach are noteworthy. Firstly, students by virtue of their ranked position got a normative indication of their performance relative to the entire cohort. Additionally, there was an immediate context for this performance as students evaluated a number of their peers work. The immediate feedback was therefore both situational and contextual. Secondly, the focus on appraisal, which possibly also became more sophisticated over the four assignments, engaged students in determining the critical aspects of capability and supported the creation of individual constructs of capability. This discussion with the self was mediated by authentic evidence that exemplified varying qualities and standards. The exposure to peer work was immediate and unqualified, requiring that each student critiqued evidence as the basis of constructing or refining his or her own construct of capability. In essence this unmediated feedback acted as an explicit feed-forward.

Interestingly, the impact of the process was variable across the cohort. Students who initially performed poorly showed a significant gain over the course of the four assignments. This is also evident by their movement into higher quartiles by the end of the semester. The middle quartiles had a more general dispersal however still saw an overall increase in performance. The initially highest performing students were the only group to see an ultimate decrease in performance. Many students also ended up in lower quartiles. The capacity to move should be considered with respect to a rank order, i.e. students from the bottom of the rank initially had more potential to move up in comparison to the top performing student who could only remain in that position or move lower. However as the rank is defined by relative standards considering the mean scores could be somewhat misleading as there was a statistically significant increase in performance between assignment 3 and 4 for Q4 suggesting a new standard in relative terms. There

is some confidence in this postulation as the translation of rank to percentage was independently and externally moderated.

The significant improvement of Q1 is apparent, however if a student started in Q1 there were likely to remain there. Therefore, they had the greatest capacity to improve; yet they were never likely to improve beyond the mean improvement of the entire cohort. While there is little doubt that being exposed to superior work enhanced the performance of Q1 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. Interestingly, despite not moving quartile this increased performance was sustained across the remaining assignments.

The unarticulated nature of this ACJ and ipsative approach may have a drawback for the Q1 students. They may be able to recognise work that is better than the work they produced and as a result be able self-audit and set relative goals, targets and standards for the next assignment. However, it is also possible that they cannot conceive new standards unless they have seen evidence of what a new standard looks like and receive additional support in the formulation of such new standards into clear targets. The results would suggest that Q4 students demonstrated a different capacity. The initial exposure to peers work seemingly affirmed their position as top quartile students. This possibly resulted in Q4 students believing that they understood the standard and assumed that the standard was static. This hypothesis is supported by the mean reduction (−10.97) in performance in Assignment 2. Furthermore, a possible recalibration of standards resulted in a non-significant improvement between Assignment 2 and Assignment 3 (MD=.130). Similar to Q4 student's ability to define the standard in Assignment 1, once the rank was established to be dynamic and improving, they again appear to have demonstrated their ability to set a new standard and record a statistically significant improvement in Assignment 4. This could be considered a new standard and the discriminating factor between the capability of Q1 and Q4 performance.

Finally, in relation to the time taken to complete judgments across each of the assignments, the longest mean time was observed in Assignment 4 (259.770 s). From a pragmatic perspective, considering the average number of judgments that students made in each ACJ session was between 8 and 9 judgments, this approach to assessment and feedback is quite feasible in educational settings. Interestingly, a trend is observed that the average time taken to make judgments increased with each assignment. It could be posited that as the students constructs of capability developed over the course of the assignments, the time needed to make a judgment should decrease as students would become more efficient in identifying the quality of work. This theory would align with the performance results in this study whereby student performance increased over each assignment. Therefore, it is speculated that as students' constructs of capability may have evolved, the responses to each assignment became more sophisticated and as a result more time was needed to make judgments. This increased level of sophistication reinforces the conjecture that although Q4 students had a mean decrease in performance between Assignment 1 and Assignment 4, the quality of their work had improved, reinforcing the position that the mean performance scores do not in themselves provide sufficient relative information from an ipsative perspective to assessment.

References

Argyris, C. (1976). Single-loop and double-loop models in research on decision making. *Administrative Science Quarterly, 21*(3), 363–375.

Baker, E., Ayers, P., O'Neill, H., Choi, K., Sawyer, W., Sylvester, R., et al. (2008). *KS3 English test marker study in Australia: Final report to the National Assessment Agency of England*. London: QCA.

Bartholomew, S., Nadelson, L., Goodridge, W., & Reeve, E. (2018a). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment*. https://doi.org/10.1080/10627197.2018.1444986.

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

Bartholomew, S., Strimel, G., & Jackson, A. (2018b). 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., Strimel, G., & Yoshikawa, E. (2018c). 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., & Yoshikawa-Ruesch, E. (2018). A systematic review of research around adaptive comparative judgement (ACJ) in K-16 education. In J. Wells (Ed.), *CTETE—Research monograph series* (Vol. 1, pp. 6–28). Virginia: Council on Technology and Engineering Teacher Education.

Betts, L., Elder, T., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment & Evaluation in Higher Education, 34*(1), 1–15.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. London: Routledge.

Buckley, J., Seery, N., Power, J., & Phelan, J. (2018). The importance of supporting technological knowledge in post-primary education: A cohort study. *Research in Science & Technological Education*. https://doi.org/10.1080/02635143.2018.1463981.

Canty, D., Seery, N., & Phelan, P. (2012). Democratic consensus on student defined assessment criteria as a catalyst for learning in technology teacher education. In T. Ginner, J. Hallström, & M. Hultén (Eds.), *PATT2012: Technology education in the 21st century* (pp. 119–125). Stockholm: PATT.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407.

Cattell, R. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303.

Coates, H. (Ed.). (2014). *Higher education learning outcomes assessment—International perspectives*. Frankfurt: Peter Lang.

de Vries, M. (2016). *Teaching about technology: An introduction to the philosophy of technology for non-philosophers*. Basel: Springer.

Hager, P., & Butler, J. (1996). Two models of educational assessment. *Assessment & Evaluation in Higher Education, 21*(4), 367–378.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Holmes, N. (2014). Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment & Evaluation in Higher Education, 40*(1), 1–14.

Hughes, G. (2011). Towards a personal best: A case for introducing ipsative assessment in higher education. *Studies in Higher Education, 36*(3), 353–367.

Hughes, G., Wood, E., & Kitagawa, K. (2014). Use of self-referential (ipsative) feedback to motivate and guide distance learners. *Open Learning: The Journal of Open, Distance and E-Learning, 29*(1), 31–44.

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment & Evaluation in Higher Education, 37*(1), 125–136.

Kaufman, J., Baer, J., Cropley, D., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts, 7*(4), 332–340.

Kimbell, R. (2009). Holism and the challenge of teachers' judgement. *Design and Technology Education: An International Journal, 14*(1), 5–6.

Kimbell, R. (2012). The origins and underpinning principles of e-scape. *International Journal of Technology and Design Education, 22*(2), 123–134.

Kimbell, R., Stables, K., Wheeler, A., Wosniak, A., & Kelly, V. (1991). *The assessment of performance in design and technology*. London: Schools Examinations and Assessment Council/Central Office of Information.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009). *E-scape portfolio assessment: Phase 3 report*. London: Goldsmiths College.

Lea, M., & Street, B. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157–172.

McDermott, P., Fantuzzo, J., & Glutting, J. (1990). Just say no to subtest analysis: A critique on wechsler theory and practise. *Journal of Psychoeducational Assessment, 8*(3), 290–302.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy and Practice, 21*(2), 205–220.

Newman, G. (2017). How we know, what we should know: The building blocks of cultural awareness in design education. In E. Norman & K. Baynes (Eds.), *Design epistemology and curriculum planning* (pp. 28–31). Leicestershire: Loughborough Design Press.

Nicol, D. (2007). Principles of good assessment and feedback: Theory and practice. In R. Harris & A. Muirhead (Eds.), *Proceedings of the REAP international online conference on assessment design for learner responsibility* (pp. 1–9). Glasgow: REAP.

Nicol, D., & Macfarlane-Dick, D. (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), 157–170.

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

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*(3), 277–289.

Reeves, T. (2006). How do you know they are learning? The importance of alignment in higher education. *International Journal of Learning Technology, 2*(4), 294–309.

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301–315.

Robinson, R. (2001). Calibrated peer review™: An application to increase student reading & writing skills. *The American Biology Teacher, 63*(7), 474–480.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 45–63). Dordrecht: Springer.

Scoles, J., Huxham, M., & McArthur, J. (2013). No longer exempt from good practice: Using exemplars to close the feedback gap for exams. *Assessment & Evaluation in Higher Education, 38*(6), 631–645.

Seery, N. (2017). Modelling as a form of critique. In P. J. Williams & K. Stables (Eds.), *Critique in design and technology education* (pp. 255–273). Singapore: Springer.

Seery, N. & Buckley, J. (2016). The validity and reliability of adaptive comparative judgements in the assessment of graphical capability. In J. Birchman (Ed.), *ASEE engineering design graphics division 71st mid-year conference* (pp. 104–109). Nashua, NH: ASEE.

Seery, N., & Canty, D. (2017). Assessment and learning: The proximal and distal effects of comparative judgment. In M. de Vries (Ed.), *Handbook of technology education* (pp. 1–14). Basel: Springer.

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.

Shavelson, R. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*(2), 73–86.

Swan, M. (2006). *Collaborative learning in mathematics: A challenge to our beliefs and practices*. London: National Institute for Adult and Continuing Education (NIACE) for the National Research and Development Centre for Adult Literacy and Numeracy (NRDC).

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34*(4), 273–286.

Vaessen, B., van den Beemt, A., van de Watering, G., van Meeuwen, L., Lemmens, L., & den Brok, P. (2017). Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: A pilot study. *Assessment & Evaluation in Higher Education, 42*(6), 872–886.

Weaver, M. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education, 31*(3), 379–394.

Zimmerman, B. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17.

Zlatkin-Troitschanskaia, O., Pant, H. A., & Coates, H. (2016). Assessing student learning outcomes in higher education: Challenges and international perspectives. *Assessment & Evaluation in Higher Education, 41*(5), 655–661.

Zlatkin-Troitschanskaia, O., Shavelson, R., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411.