Sharing and securing learners' performance standards across schools

Emeritus Professor Richard Kimbell
Goldsmiths University of London, United Kingdom

https://doi.org/10.37517/978-1-74286-685-7-6

Richard Kimbell's specialist interest is in design learning and its interaction with assessment. He founded the Technology Education Research Unit at Goldsmiths University of London in 1990, and for over 25 years, ran research projects for research councils, for industry, for government departments, as well as for professional and charitable organisations. Richard has published widely in the field of technology education, including five single-authored books, has written and presented television programs and regularly lectures internationally. He has been a consultant to the National Academy of Engineering and the National Science Foundation in the United States, and a visiting professor at the University of British Columbia, the University of Stockholm, Edith Cowan University in Perth, Australia, Texas A&M University, and The University of the Shannon, Ireland.

Abstract

Assessing learners' performance makes very different demands upon teachers depending on the purpose and the context of the assessment. But common to all assessment is some sense of what 'quality' looks like. Most often teachers engage in formative assessments in the classroom, and the familiar standards of the classroom are adequate for this purpose. However if teachers are to undertake external, nationally regulated assessment then some sense of a national standard of quality is required. But there are very limited mechanisms by which teachers can acquire this understanding, so they use their best judgement, and standards vary from school to school not because anyone is attempting to cheat the system but simply because they cannot know what the real national standard is. It is for this reason that regulated examination bodies follow some process such as the following from the State Examinations Commission (SEC) in Ireland. '… teacher estimated marks will be subjected to an in-school alignment process and later a national standardisation process'. (SEC, 2021). How much simpler it would all be if teachers had – as a matter of normal practice – access to, and familiarity with, work from a national sample of schools, not just their own classroom.

Adaptive Comparative Judgement (ACJ) is an online assessment tool that has been used for some years, principally as a formative tool for learners (e.g. Bartholomew et al., 2018; 2019). This presentation reports on a study of the new ACJ Steady State tool from the same stable. The purpose of the new tool is to solve the problem of variable standards across schools by enabling teachers to make paired judgements of work from multiple schools and thereby evolve and agree standards of performance beyond their own school. The current study is operating in Ireland with a group of schools, a university, and the SEC. The anticipated outcomes include 1) better consistency of performance standards across schools in the research group and 2) greater understanding of and confidence in assessment judgements by the teachers. If ACJ has proved to be a powerful formative assessment tool for learners, ACJ Steady State is designed to be a formative assessment tool for teachers, helping to inform and support their assessment judgements.

Introduction

Teachers are constantly engaged in assessment activities, principally to better understand their learners' current performance. Mostly these assessment activities are informal and based around the on-going classroom tasks being undertaken by learners. Occasionally they have a more formal status, for example, when a teacher is asked to predict (by the school or perhaps by a parent) how well the learner is likely to perform in an examination. And even more occasionally, the teacher is asked to make a clear formal assessment of the standard of performance achieved by a learner as part of an external examination. As the teacher's judgements progressively migrate from informal-classroom to external-examination, they involve an ever-closer association with standards of performance outside the school; with those of other schools in the town; or in the next county; or even nationally. The problem for the teacher is that there is almost no mechanism by which they can know what those standards are. Examination bodies publish lists of assessment criteria and sometimes these are associated with examplar materials, but every possibility cannot be covered and – in any event – quality cannot be defined in words (Polanyi, 1958). So inevitably the standards remain mysterious and open to interpretation and even misunderstanding.

While performance standards *within* a school can be carefully monitored and adopted by a teacher, it is almost impossible for that teacher to do the same thing *across* schools. This difficulty inevitably results in assessment error (The Office of Qualifications and Examinations Regulation [Ofqual], 2021); in teachers awarding different 'scores' for the same quality of work simply because of the lack of familiarity with the standards of work that apply in this school or that school and by the differences in the standards 'held' by the teachers as yardsticks of performance.

> In simple terms, error is the difference between the result that a student ought to have been awarded from an assessment – given their level of attainment in the subject being assessed – and the result that they ended up being awarded (Ofqual, May 2021).

Comparative judgement for assessment

Between 2005 and 2010 at Goldsmiths, we ran project e-scape, a project commissioned by the Qualifications and Curriculum Authority (QCA) in England. The purpose was to explore an approach to performance assessment based on portfolios that could be created digitally by learners in the classroom and assessed digitally (online) by teachers and others. As part of that project, we created a new assessment tool based on the idea of comparative judgement (see Kimbell et al., 2009; Williams & Kimbell, 2012). The new tool, ACJ, was based on work by Laming (2004) who argued that all human judgements are comparative, and by Pollitt (2004) who first used comparative judgement in reliability studies for the University of Cambridge Local Examination Syndicate, a forerunner of Cambridge Assessment. The ACJ tool was used in prototype form in a pilot study in 2008 and as a developed tool in a national study across England in 2009. The results were encouraging, as Pollitt reported in the final e-scape report in 2009.

> The final scale spread the portfolios out with a standard deviation of almost 3 units. The average measurement uncertainty for a portfolio was about 0.67 units, and the ratio of these two figures was 4.45. This means that the standard unit of the scale was almost 4.5 times as large as the uncertainty of measurement. This means the portfolios were measured with an uncertainty that is very small compared to the scale as a whole; this ratio is then converted into the traditional reliability statistic – a version of Cronbach's alpha or the KR 20 coefficient. The value obtained was 0.95, which is very high in GCSE terms (Kimbell et al., 2009, pp.73–81).

Quite apart from the reliability issue, however, what struck us most forcibly during the assessment process were the responses of the teachers undertaking the assessments. There were several hundred portfolios, all available online, and the ACJ system presented two at a time, inviting the teacher/judge to decide which was the stronger of the two. While the judge training involved the identification of headline criteria, these were not to be separated and scored but rather to be 'held-in-mind' as the judge came to a single holistic decision. Which is stronger … this one or that one? All the teachers had supervised the e-scape classroom activity in their various schools and were familiar with the work. But – for the first time – they were able to see learners' work not just from their own school but also from schools all over the country. They were fascinated to see how different approaches to the work had developed and how their own students' performances related to others.

The ACJ tool has now become commonplace in schools, principally as a formative assessment tool. Asking learners themselves to make the paired judgements has the effect of promoting discussion of what 'quality' means in a piece of work, and teachers are adept at using such discussions to help enhance learners' performance. See for example the extensive collection of work by Bartholomew and colleagues in Utah in the United States (Bartholomew et al., 2018; 2019).

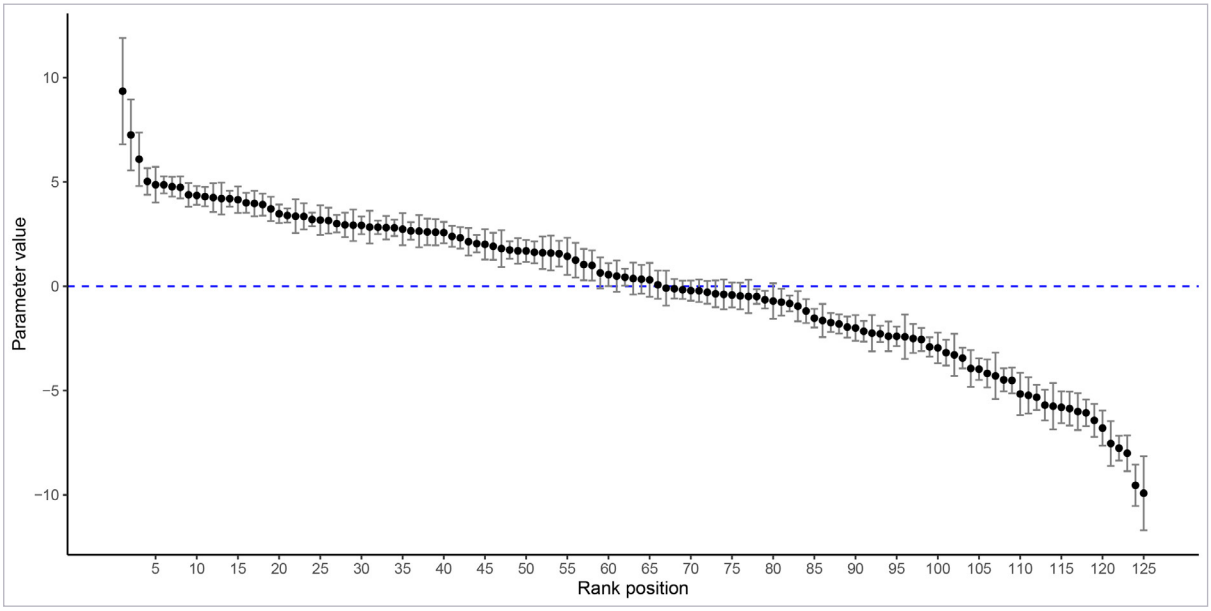
But, up to now, there has been a limitation with ACJ in that it is a single-cohort tool. So a class in School A can do an ACJ exercise, and another class in School B can do a different one. This will result in two performance ranks, but with no means to relate them. Of course the two schools could collaborate on a single bigger ACJ session, but that is much more complex to arrange and to manage for a class teacher. We came to see that it would be very useful if separate ACJ ranks could – in some way – be combined into a single rank. Thus was born the idea of a new form of ACJ that we have provisionally called 'ACJ Steady State'. This is being developed by RM Education (a leading supplier of learning and assessment resources to the education sector) who acquired and refined the original ACJ algorithm. (see https://www.rm.com/). In collaboration with RM, a pilot study is underway in Ireland involving the SEC and the Technological University of the Shannon. The study involves a trial with teachers and learners in 10 schools pursuing a graphics program at Leaving Certificate level.

A pilot study of ACJ SteadyState

If we imagine that a class of learners in four schools A,B,C,and D has each produced (say) 25 pieces of work, we can collect a small team of teachers to undertake a standard ACJ process on the 100 portfolios. This will result in a simple rank order of the 100 pieces.

For the purposes of this study, this rank then becomes 'the ruler', against which we might choose to measure other work.

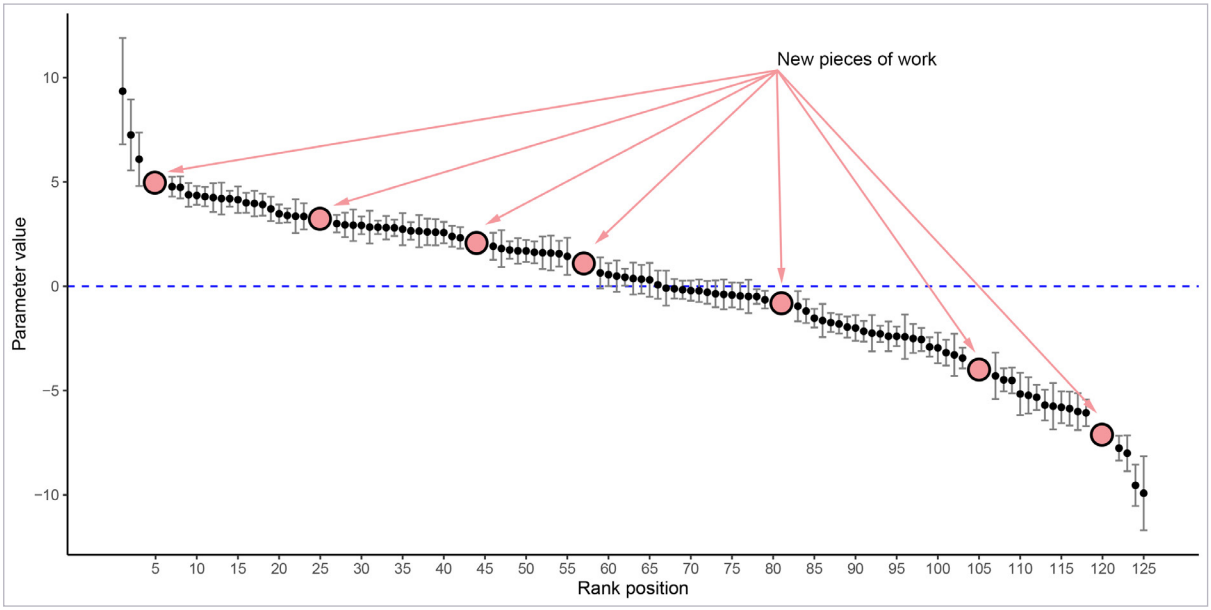
Figure 1 ACJ rank becomes 'the ruler'



If we also now imagine a fifth school (E) that has also undertaken the same exercise as the classes in schools A–D. This new work (another 25 pieces) will be added to the pot of portfolios and will be judged by the same team of teachers, but this time using ACJ Steady State.

It is important to recognise that in normal ACJ exercises, the positions of *both* of the portfolios will change as the result of a judgement. The rank continuously modifies itself as the judging unfolds. But ACJ Steady State operates differently. The ruler is fixed and the judging thereafter only adjusts the position of the new pieces until they reach their points of stability within the fixed rank.

Figure 2 ACJ Steady State with new work located in the ruler



Provided this all works as we believe that it will, then it will be possible to create a ruler with a selected number of schools, and subsequently we will be able to integrate new schools into the ruler. All it requires is that the judges make their judgements of the new school's work against the work in the ruler. ACJ Steady State is, in effect, a multi-cohort tool and we can go on and on adding schools. It will thus be possible to create not just school ranks but also regional ranks and even a national rank. It should be noted that the judging process for ACJ Steady State will be lean and quick. The algorithm will only select pairs involving a new portfolio, sometimes comparing it to a 'ruler' portfolio and sometimes to another new one. Moreover the new algorithm only has to calculate and adjust the position of that new piece. While the ruler itself (within ACJ) required approx 15 rounds of judging to fix the positions of the 100 pieces of work, our modeling suggests that (within ACJ Steady State) five or six judgements will fix the positions of the new work.

Issues to be explored

The issues to be tackled here include some that might be thought of as informal, formative issues for teachers and schools, while others concern more technical matters that have to be dealt with for formal assessments.

Formative classroom issues

1. It is critical to the purpose of the project that the teachers are empowered to discuss the judgements they make. Wherever possible teachers will work as pairs, agreeing the judgements as they go and noting the qualities of performance that led to the judgements. We believe that it will be possible for a group of teachers not just to operate reliably but also to articulate and agree as a group what the *qualities of work* are at all levels through the rank.
2. We want to explore the pedagogical implications of the ruler. When teachers are self-consciously aware of the standards of performance that make up the scale, what do they choose to do with that information? Might they choose to use ACJ Steady State as a regular progress check for themselves? Might they engage learners in using it to observe work beyond their own school? Might they choose to create rulers not just for overall performance but also for *elements* of performance?

Technical assessment issues

1. When schools (or the SEC) wish to create a ruler that reflects a 'national standard' it is important to consider how this might be done. Should this be from a random group of schools? How big should the ruler be? Should it include selected work from a previous year's performance? What happens if work from a new school goes off the scale at one of the ends? The project will enable us to explore and recommend a model of practice for deriving and applying the ruler.
2. Once a new school has been integrated into the ruler, it will be possible not only to comment on the placing of individual portfolios, but also to derive a 'school statistic' that identifies how, for example, the mean or median performance of School A is different from the mean or median performance in the ruler. This might enable a moderation tool to be developed to assist the SEC in their desired adjustment of teacher assessments 'teacher estimated marks will be subjected to an in-school alignment process and later a national standardisation process' (SEC, 2021).

Conclusion

Currently it is only at external examination times that teachers get the opportunity to see their learners' work set against a backdrop other than their own school's, and at that moment it is already too late for the information to feed into positive learning experiences. The existing ACJ tool is currently in use in classrooms where teachers use it as a formative feedback tool to help learners to get a better understanding of their own work. Our vision for ACJ Steady State is that it should be *a formative feedback tool for teachers*; enabling them to share work across multiple classrooms and debate standards *across* schools. We see this as a tool that will help to de-mystify national standards, helping schools to collaborate and allowing teachers to gain confidence through their shared judgements.

References

Bartholomew, S. R., Strimel, G. J., Garcia Bravo, E., Zhang, L., & Yoshikawa, E. (2018). *Formative feedback for improved student performance through adaptive comparative judgment* [Paper presentation]. 125th ASEE Conference, Salt Lake City, Utah, United States.

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education, 29*(2), 363–385.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Davies, D., Martin, F., Pollitt, A., Whitehouse, G.. (2009). *E-scape portfolio assessment: a research and development project for the Department for Education & Skills (DfES) and the Qualifications and Curriculum Authority (QCA), phase 3 report*. Technology Education Research Unit, Goldsmiths, University of London.

Laming, D. (2004). *Human judgement: The eye of the beholder.* Thomson.

The Office of Qualifications and Examinations Regulation (Ofqual), (2021, 17 May). Bias in teacher assessment results. *The Ofqual Blog* https://ofqual.blog.gov.uk/2021/05/17/bias-in-teacher-assessment-results/

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy.* Routledge & Kegan Paul Limited.

Pollitt, A. (2004, September). *Let's stop marking exams*. [Paper presentation]. IAEA Conference, Philadelphia, United States. https://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf

State Examinations Commission [Ireland] (2021) *Leaving certificate 2021 and accredited grades*. https://www.citizensinformation.ie/en/education/state_examinations/leaving_certificate_2020_calculated_grades.html

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