frontiers in Education

PERSPECTIVE
published: 18 March 2022
doi: 10.3389/feduc.2022.806894

Using Teachers' Judgments of Quality to Establish Performance Standards in Technology Education Across Schools, Communities, and Nations

Niall Seery[1], Richard Kimbell[2] and Jeffrey Buckley[1,3]*

[1] Faculty of Engineering and Informatics, Technological University of the Shannon, Athlone, Ireland, [2] Goldsmiths, University of London, London, United Kingdom, [3] Department of Learning, KTH Royal Institute of Technology, Stockholm, Sweden

The establishment and maintenance of national examination standards remains a serious issue for teachers and learners, whilst the levers of control remain firmly in the hands of Awarding Bodies and supervising politicians. Significantly, holistic assessment presents an agility and collective approach to establishing in the minds of teachers "what is of value" when determining the comparative evidence of pupil performance. It is argued in this paper that the collation of the comparative judgment process can initially identify and subsequently maintain standards of performance that can be defined on a cluster, regional or even national level. Much comparative judgment research centers on the formative benefits for learners, but here we place the focus on teachers operating in collaborative groups to establish standards within and beyond their own schools, and ultimately across the nation. We model a proof-of-concept research project. A rank is produced by the collective consensus of the participating teachers and used to simulate a definition of standard. Extrapolations are statistically modeled to demonstrate the potential for this approach to establishing a robust definition of national standards. But central to the process is what is going on in the minds of teachers as they make their judgements of quality. The research aims to draw out teachers' constructs of quality; to make them explicit; to share them across classrooms and schools; and to empower teachers to debate and agree their standards across schools. This research brings to the fore the symbiotic relationship between teaching, learning and assessment.

Keywords: national standards, teacher judgment, comparative judgment, validity, assessment

OPEN ACCESS

Edited by:
Tine Van Daal,
University of Antwerp, Belgium

Reviewed by:
Wei Shin Leong,
Ministry of Education, Singapore

****Correspondence:***
Niall Seery
nseery@ait.ie

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *01 November 2021*
Accepted: *02 February 2022*
Published: *18 March 2022*

Citation:
Seery N, Kimbell R and Buckley J (2022) Using Teachers' Judgments of Quality to Establish Performance Standards in Technology Education Across Schools, Communities, and Nations. Front. Educ. 7:806894. doi: 10.3389/feduc.2022.806894

INTRODUCTION

Much of the focus of Adaptive Comparative Judgment (ACJ) research is centered on cohort-based application cases (Williams and Kimbell, 2012; Bartholomew and Jones, 2021), where the agendas include formative (e.g., Bartholomew et al., 2019) and summative (e.g., Jones and Alcock, 2012; Whitehouse and Pollitt, 2012) application, sometimes combining both formative and summative agendas to frame an assessment "as" learning approach (Seery et al., 2012). Studies report high

levels of reliability (Bartholomew and Yoshikawa-Ruesch, 2018; Bartholomew and Jones, 2021), and this gives confidence in the rank order produced by the binary judging session. ACJ uses an adaptive algorithm to govern the presentation of pairs of student "portfolios" which are then holistically compared by a cohort of "judges," e.g., teachers, on evidence of learning. Research by the Assessment of Performance Unit (APU) team at Goldsmiths in the 1980s empirically demonstrated that teachers were far more reliable when comparatively assessing whole pieces of work than they were when assessing individual qualities. When assessing writing performance through a comparative approach, teacher judgment is reported as "highly internally consistent when judging quality" (Heldsinger and Humphry, 2010, p. 221). These binary decisions on authentic evidence ultimately position students' work on a rank from top to bottom, as described by Pollitt (2012a,b). The approach produces associated parameter values or "ability scores" which are indications of relative differences between portfolios along the rank.

Considering the context of Design and Technology, the creative relationship between designing and making is difficult to reflect in a criterion driven assessment and this approach can in fact change the very nature of the activity to conform to what is weighted as valued output. The challenge in developing a retrospective portfolio and even an artifact is influenced by the criterion specified before the task begins. The work of the Kimbell et al. (1991) established the need to consider student work differently and framed the importance of a holistic view of performance. Like consensual assessment techniques (CAT; Amabile, 1982), the aggregation of expert judgements through an ACJ approach provides a reliable and valid approach to measuring (Bramley, 2015; Bramley and Wheadon, 2015; Coertjens et al., 2017; Verhavert et al., 2018; Kimbell, 2021). Aligned with the approach of using calibrated exemplars (Heldsinger and Humphry, 2013), this research proposes a "bottom up" development of national standards.

RESEARCH AGENDA

Up to this point, ACJ has been used with groups of learners with the purpose of arriving at a performance rank of those learners for formative and/or summative purposes. If 100 learners are participating in an ACJ session, then, using ACJ, judges can arrive at a performance rank for those 100 learners. But the practice becomes more complex if large cohorts are anticipated, as would be the case for national examinations. Annually in Ireland, approximately 60,000 students take examinations as part of the Leaving Certificate—a State organized national examination taken at the end of secondary level education with results feeding into a matriculation system for tertiary education admissions. Managing such a number through an ACJ exercise would be extremely challenging. However, it is not the purpose of this project to attempt such an exercise. Rather we seek to investigate the use of a new form of ACJ to begin to explore what would be involved in building a system that enables teachers to collaborate across schools to arrive at a view of a national performance standard. Broadly speaking this system would start with a locally established standard (within a school or small cluster of schools) and move progressively to regional groupings of schools (e.g., across cities/counties) and ultimately to a view of a standard across the entire nation.

To begin this inquiry, three initial steps are required to first establish a performance standard at a local level:

1. First, standards will be tentatively defined and agreed between teachers in collaborating schools in terms of learners' performance on an ecologically valid activity.
2. Second, the work will be combined into a single ACJ session which will be judged by the teachers who had supervised the work in the collaborating schools. This is a classroom-based view of standards in which teachers do not seek to apply a standard devised by someone else, but rather they will create a rank of the work using holistic judgments guided by their own personal constructs of capability which can then be used to refine and clarify standards.
3. Third, through a collaborative process, teachers will reflect upon the ACJ process and refine the initially determined standards and then map these back onto the rank order produced through the ACJ session to create a "reference scale" of work which other, new pieces of work can be positioned along.

This research agenda is to establish a reference scale that can be used to position students' work relative to a national standard, built from an initial local but representative standard, and in essence determine national standards through authentic performance. It is useful to think of this "Steady State" as a type of *Ruler*. Importantly, the Ruler would be produced from authentic evidence of student work, brought about by the judgments of teachers making decisions on authentic work. This Ruler reference would move away from abstracted criteria or the need for interpretation, instead the real evidence of learners' work would form the basis of comparators, representing the quality of work presented. The idea is that this approach would improve standards by improving the whole performance and supporting the developed conceptualization of what constitutes quality.

Central to the establishment of the Ruler is drawing from teachers "what is of value" when judging student work. This research centers on using the holistic approach of ACJ to establish a reference rank that is built on teacher expertise. Jones et al. (2015) argue that freeing judges from predefined criteria can enable judges to tap into their expertise. This view is supported by van Daal et al. (2019) and is seen as particularly useful when assessing complex competencies (Pollitt, 2012a). Comparative methods enable the teacher to obtain reliable scores for complex skills more easily than using analytic methods (Jones and Inglis, 2015; Coertjens et al., 2017). Barkaoui (2011) also found that holistic marking favored higher levels of consistency between markers in comparison with analytic marking. The work of Lesterhuis et al. (2018) is also relevant as their findings report that the considered construct of quality in a comparative method is multidimensional and notes that teachers use their experience to give meaning to the evidence. Not only can the "bottom

up" approach to standards setting unleash the expertise of teachers, but building on the insights shared by Van Gasse et al. (2019) where they highlight the change in the conception of the assessor following a CJ intervention, it can also support the development of the teacher in terms of refining their constructs of capability. The richness of varying perspectives discussing the concrete context of the assignment, ensures task-orientated and not examiner-orientated focus. As highlighted by Heldsinger and Humphry (2010), the potential of using a calibrated sale of exemplars to compare future work is the technical focus of this proposed work.

MODELING AN APPROACH TO DEFINING NATIONAL STANDARDS

The planned scoping project will operate through three phases, and the subject of study will be "Design and Communication Graphics" which has an annual Higher Level examination cohort of circa 4,000 students. To get a representative sample to define a national sample, the following approach and calculations have been produced.

- A total population of 3,804 is the subject of this proof-of-concept research cohort and a sample of 250 portfolios was calculated to give a confidence level of 95% with a margin of error of 5.99%.
- A selection of judges will participate in a normal ACJ session with the work. Judging sub-groups will be identified to include teachers, researchers, and experienced examiners. The representative rank order will be produced by the judges making binary decisions on a combined pool of portfolios from all over the country, so the teachers are not merely judging their own learners but are exposed to a representative sample of evidence. This will produce an agreed performance rank for the 250 learners based on the judgments of the teachers, researchers, and examiners. Reliability statistics will be carefully examined to explore any differential effects of the teachers' judgments. The performance rank will thereafter be called The Ruler.
- An additional sample of 50 portfolios will also be randomly selected from the population data that are independent from the work that created the Ruler. At the end of the project, the resulting 50 pieces of work will be judged against the Ruler using a modified ACJ tool. The purpose is to (1) explore approaches to positioning these along the Ruler and then (2) to see where and how the 50 new pieces locate themselves along the Ruler. The judgments will be made by the original team of teachers, researchers, and examiners.
- The research questions will focus on the length and precision of the Ruler by exploring teacher judgment and the variations in the ways things can be valued. The process of consensus building on what is of value within the process of building the ruler and ensuring the bias management and representation test the lens appropriate to its utility as an instrument for national standard definition.

The agenda to establish a means by which a national standard can be determined from the evidence produced by pupils and adjudicated on by in-practice teachers, highlights several research considerations discussed in the following section.

GAP ANALYSIS

Extending the application of the ACJ process beyond that of a single cohort or cluster to form a national picture of performance brings into focus the details of the ACJ reliability statistics, parameter values and the association with validity (cf. Buckley et al., 2022) all of which are of particular interest to this research. Although reported reliability statistics of more than 0.9 give confidence (Pollitt, 2012b; Bartholomew and Jones, 2021), there are notable critiques of the adaptive process. Bramley has identified that the adaptive algorithm artificially inflates estimates of the reliability of the outputted rank order of work (Bramley, 2015; Bramley and Wheadon, 2015). Much of the issue is caused by a "spurious separation among the scripts" (Bramley, 2015, p. 14) where the adaptive algorithm makes it impossible, for example, for work that "loses" a small number (e.g., two) of judgments against work when paired truly at random to show that it is actually relatively good work as the adaptive algorithm will make it less likely to get paired with work that won in those initial rounds. Further, the process of ACJ has issues at the extremes. To take the piece of work that "wins" or is ultimately placed at the top of the rank, it is likely that it may never or will rarely ever be judged as a losing piece of work in a pairwise comparison. As such, there is little information about the work compared to those determined as closer to average. The winning piece of work may confidently be positioned at the top, but there is much uncertainty regarding its parameter value. While these issues do not affect the absolute rank, they do affect the validity of interpreting and using the parameter values as denotations of relative distances between pieces of work, which is problematic when we seek to develop the application beyond a single cohort.

A number of studies have aimed to address these problems. First, Bramley and Vitello (2019) note some potential advantages of adaptivity. These included that adaptivity can increase efficiency by avoiding pairing portfolios which are very far apart on the rank, and on the issue of inflated reliability they note that while adaptivity may spuriously inflate the standard deviation, it could actually reduce error. One possible approach to addressing inflated reliability which will be explored in this project is to increase the number of comparisons. Verhavert et al. (2019) note that in CJ, to reach a reliability of 0.90 without adaptivity, 26–37 comparisons are needed per portfolio. Bramley and Vitello (2019) point out that the reason for the inflated SD is that the introduction of adaptivity means most portfolios would be compared indirectly via other portfolios. Therefore, it is possible that the use of adaptivity to select the portfolios for comparison which would provide the most information with a minimum number of comparisons, such as 26–37 per portfolio, used as a stopping rule as opposed to the use of a reliability threshold could provide a suitable solution. Such a minimum number will need to

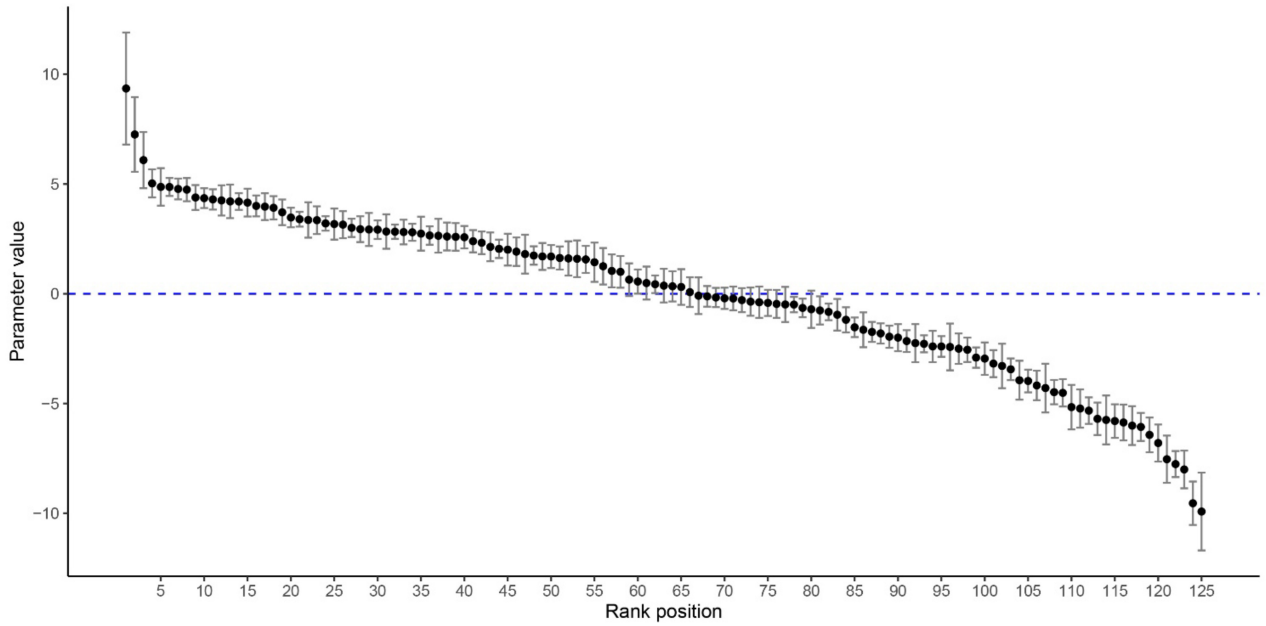

FIGURE 1 | The goal of step 1 is to develop a normative rank order of work, which includes information regarding locally derived standards inductively generated from reflection on authentic work. Note that the above figure consists of simulated data from an arbitrarily defined 125 pieces of work for illustrative purposes only.

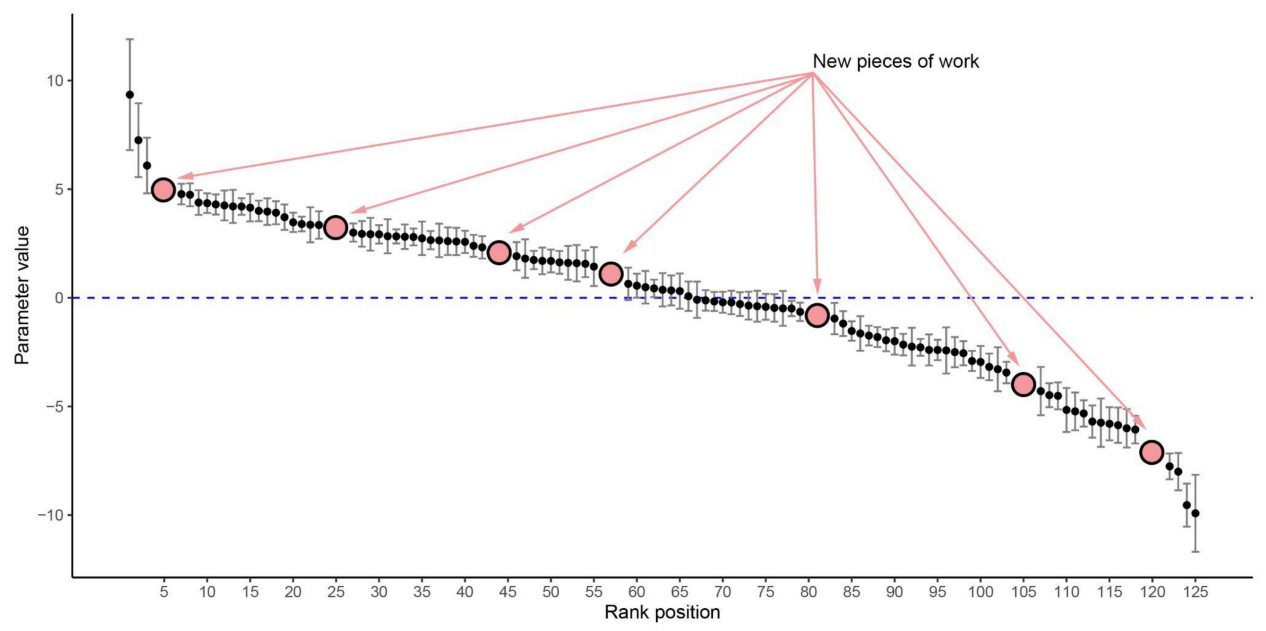

FIGURE 2 | The goal of step 2 is to use the Ruler, with a new ACJ approach, to position new pieces of work against, illustrated hypothetically in the above figure.

be determined, and in doing so reliability deflation (Bramley and Vitello, 2019) should be considered.

The research therefore will involve a two-stage process. First creating the Ruler (through a normal ACJ judgment process—**Figure 1**) and second employing the Ruler in a separate judgment process to seek to locate new pieces of work within the quality scale defined by the Ruler (**Figure 2**). This second process clearly requires a different ACJ approach. In "normal" ACJ judging sessions, all the work is floating and is affected by each judgment. A new comparison judgment will therefore affect the position of both pieces of work involved in the comparison. What we are proposing is that once the Ruler has been agreed it is fixed, and judgments thereafter (in the 2nd phase of judging) are intended simply to locate the new pieces within the Ruler.

The standards articulation process can occur at several points. It can be inserted at the end of the first judging round and lead to an articulation by teachers of the Ruler, but then it can occur again after the second phase of judging locates the new work into the Ruler. Understanding these standards is at the heart of this project and teachers will become very familiar with iterative discussions that seek to clarify and refine them. It is our belief that teachers have a working understanding of quality standards that enable them to distinguish good from mediocre work and mediocre from poor performance. But these standards are typically internal to their practice. Our aim is to draw them out through a process that (1) requires the teachers to use them in a judging round, and (2) through discussion empowers teachers to articulate what indications and qualities they see in the work that makes them judge it as outstanding/good/adequate/poor. Whitehouse and Pollitt (2012, p. 15) highlight that "thought needs to be given to how shared criteria can be exemplified and disseminated." At the end of the second phase a range of statistical exercises can be undertaken with the resulting data. One might judge, for example, that the distribution of the new pieces of work from the second pair of schools was loaded more toward the upper end of the Ruler. This can be calculated exactly. The Ruler does not merely show individual placings but can also reveal school-based performance.

This process can go on as often as required with school groups of work being endlessly judged against the Ruler, producing individual positions for the work and school–based data from the amalgamation of those placings. All schools in a region (and even across the nation) can therefore be assessed with an ACJ-judgment-style of assessment against a common standard—the Ruler. This does not require an enormous "once-for-all" ACJ exercise simultaneously involving thousands of learners. Rather it can be done in two steps by (1) establishing the Ruler and (2) subsequently comparing learners from other schools to the Ruler. The concept of a ruler affords the utility of an instrument that can order authentic work on a scale representative of the breath of performance. The disparate parameter values record the separation between units of work, enabling transposition of the rank normatively or relatively, depending on the sensitivity of the assessment context. Like previous research, ACJ can also record the judge's statistical alignment, unpacking further consensus and misfit.

Building on comparative judgment, teacher assessment and professional development in various subject contexts, this paper proposes a study that will endeavor to answer questions that focus on 3 thematic areas: considerations for teachers and schools, technological developments, and standards and awards, with the details being unpacked in the following sections.

Teachers and Schools

- Can teachers use authentic data (with ACJ judging) to articulate what standards are?
- Can teachers fully articulate what distinguishes "good" from "less good" work?
- Can teachers' decisions reach consensus and be aggregated so as to determine what is of value when considering evidence of learning?

Technological Developments

- Can we establish a valid and reliable definition of standards and thereby create a Ruler that is long enough and precise enough to cater for all performance levels?
- From a technological perspective, how can "new" work be judged into the Ruler?

Awards and Standards

- Can teachers distinguish statistically discrete levels of performance from within the Ruler?
- Can teachers use the Ruler to effectively compare other work to a National Standard?

Supplementary research questions that will be explored in parallel and not as part of the modeling study include:

- What is the impact of exposing teachers to a breadth of work from other schools on their definition of standards?
- How will this exposure impact teachers' professional development, specifically in assessment literacy?
- What is the relationship between task design and student performance?
- Do teachers' articulation of standards vary with the task?
- Are there inherent biases that impact on different categories of students?
- Can assessment tasks be designed to be independent? Or can we control task independence?
- Can we (or should we) articulate national standards as absolute and monitor performance over time?
- How could the Ruler be used in practice as a formative and pedagogical tool?

CRITICAL ISSUES FOR TEACHERS AND SCHOOLS

The first and most critical feature of this approach to standard-setting is that the standards emerge directly because of the judgments made by teachers. Whilst teachers' ACJ judgments will be informed by criteria, those criteria are not individually scored and summed. Rather, they are all "held-in-mind" to support the teacher in making an overall holistic judgment of the quality of the work. Teachers' concepts of quality, Polanyi (1958) referred to this quality as connoisseurship, are central to the approach we seek to build. Teachers discuss the strengths and weaknesses of individual pieces in a comparison, and the many finely distinguished pieces in the Ruler provide a scale that exemplifies quality at every level. Wiliam (1998) described teachers doing this in the early days of the England/Wales National Curriculum. Given pages of criteria to score, they largely ignored them, preferring to make their judgments in more holistic ways:

> . . . most summative assessments were interpreted not with respect to criteria (which are ambiguous). . . but rather by reference

> to a shared construct of quality that exists in well-defined communities of practice. (p. 6).

The work of Seery et al. (2012), exemplify the capacity of ACJ to help build quality constructs, using actual evidence as the medium for refining the emerging constructs of novice student teachers.

> A critical factor for the students was that the assessor (their peer) could empathize with their work having completed the process themselves. The process also encouraged students to engage in discussions on capability with their peers in an effort to broaden their concept and understanding of capability as the ACJ model sees judgments on students' work made across a wide range of assessors. (p. 224).

It is these constructs of quality that we shall be exploring within the community of graphic teachers. The aim will be both to build and enrich these explicit constructs and, in the process, to enable teachers to see their own learners' work against a wider frame of reference than exists in their own school. With another school . . . in another town . . . and ultimately across the nation. As teachers become more familiar with the quality of work that can be expected in relation to any task, they are empowered to develop their own practice and help their learners to improve.

CRITICAL ISSUES FOR THE "RULER"

Refining ACJ to accommodate the national standards agenda will support several critical agendas. The judging process will engage teachers in developing an understanding of what standards are, this is especially powerful when these standards are being defined by actual classroom-based evidence, where differentiation between work is established by qualities adjudicated by the teachers (van Daal et al., 2019). Building a dataset that can represent and define national standards has the capacity to build confidence in standards across schools, a bi-directional relationship where the feed-forward micro to macro definition of standards will also backwash from macro to micro to help teachers to improve the performance of their own students.

The significance of the national standard definition is critically dependent on the quality of the Ruler. Therefore, the Ruler needs to be long enough (so it captures the full range of performance, with no loss of utility at the ends) and precise enough (so work can be accurately placed on the Ruler). The statistical and technological solutions to developing a robust Ruler are apparent challenges. More nuanced are the challenges facing teachers in determining to what degree work can be distinguished into distinct units of performance and how many distinct units of performance can be measured at each grade level. Statistically, this plays out in terms of the standard deviation of the items (portfolios) and the discrimination that can be achieved by the teachers, both are critical to the reliability that can be achieved in the judging (Kimbell, 2021).

The Ruler can only be as good as the work that is imported into the algorithm. We could have a very good Ruler for the average piece of work, but a poor Ruler for excellent work. This is a key research agenda. Assuming the target will be a normal distribution, the focus of the research agenda is to ensure the precision and length of the Ruler, to cater for the full spectrum of performance. There are several approaches that could be used to test the robustness of the Ruler. We could bias the population sample or chain the judging session to "force" judging of comparisons within specific areas of the rank (at the ends for example, where usually we have the least amount of information). Using the analogy of a Microscope, the technology could be designed to have interchangeable lenses to take a focused look at categories of interest not just performance bands, but also (for example) issues of inclusion, access, and disadvantage. This perspective and approach have not ever been made manifest in earlier or even current ACJ work and is only necessary when you consider using the rank for the purposes, we intend here, for inter-school, clustered or national standard definition.

There are 3 critical issues to be considered. The ACJ algorithm and its ability to refine the information captured in relation to the spread of performance, requires critical and statistical review. That is, the length of the Ruler and the resultant graduations are sufficiently defined to represent the breadth of performance that then can be used for future comparisons. Secondly, the probability at the extremes needs to be comparable with the confidence in parameter values that emerge in the middle of the rank, with no risk of inflating the reliability of the rank. The criticism of inflation at the extremes, needs to be managed to ensure that the graduations of the analogous Ruler are consistent at the extremes and can distinguish performance effectively. Thirdly, once the Ruler is created and robustly tested the issue is to translate or transpose the rank order into a definition of standards. The creation of standards will rely on the experience of the teachers in distinguishing the discrete units of performance to form a robust dataset that represents the breadth or performance and can confidently identify grade boundaries.

CRITICAL ISSUES FOR EXAMINATION AND AWARDING BODIES

Perhaps the most critical issue for examination bodies in this approach exists in the question "How do we create the Ruler?" It would be simple enough to take a sample of schools and use that sample to create it with (say) 100 or 500 learners. But how do we ensure that it is sufficiently broadly based to capture all the levels of quality that we are concerned to identify? One possibility would be to see the Ruler as emergent and evolving, based on the standards of last year's examinations, and enriched with this year's work samples. It might therefore contain some of last year's work samples as well as this year's. This would additionally provide the possibility of a direct comparison of standards across years.

And this raises the question of the variability of performance across tasks. Examinations do not set the same questions every year. But, since they are looking to assess the same qualities, the assumption is that different questions can elicit parallel levels of performance. With task-based performance in graphics it will be interesting to see how (to what extent) parallel levels of performance can be revealed by different tasks. And critical to

that will be the articulation of the standards themselves. Teachers will initially seek to clarify their standards in relation to the task-based performance of the initiating group. But the articulation process must be sufficiently generic that it applies beyond the detail of the task itself.

There is a fine line here. The standards emerge from task-based performance, since it is the task-based performance that exemplifies those standards for the teachers to observe. But the standard needs to operate equally on parallel tasks, so it must be sufficiently specific to operate accurately on a given task but still sufficiently generic to accommodate variation. The details of this inter-task dynamic will be very revealing of teachers' views of the standard.

The paper is the starting point of a comprehensive research study that sets out to develop existing technology that has the potential to liberate teachers' professional judgment through engagement with authentic evidence of pupil learning, while establishing a definition of national standards.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

The authors received no explicit financial support for the preparation of this manuscript. Costs to support the publication of this manuscript were provided by RM Education.

REFERENCES

Amabile, T. M. (1982). Social psychology of creativity: a consensual assessment technique. *J. Pers. Soc. Psychol.* 43, 997–1013. doi: 10.1037/0022-3514.43.5.997

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assess. Educ. Princ. Policy Pract.* 18, 279–293. doi: 10.1080/0969594X.2010.526585

Bartholomew, S., and Jones, M. (2021). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *Int. J. Technol. Des. Educ*. doi: 10.1007/s10798-020-09642-6

Bartholomew, S., Strimel, G., and Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *Int. J. Technol. Des. Educ*. 29, 363–385. doi: 10.1007/s10798-018-9442-7

Bartholomew, S., and Yoshikawa-Ruesch, E. (2018). "A systematic review of research around adaptive comparative judgement (ACJ) in K-16 education," in *CTETE - Research Monograph Series*, ed. J. Wells (Reston, VA: Council on Technology and Engineering Teacher Education), 6–28. doi: 10.21061/ctete-rms.v1.c.1

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment*. Cambridge: Cambridge Assessment.

Bramley, T., and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 43–58. doi: 10.1080/0969594X.2017.1418734

Bramley, T., and Wheadon, C. (2015). "The reliability of adaptive comparative judgment," in *Paper Presented at the AEA–Europe Annual Conference*, Glasgow, 7–9.

Buckley, J., Seery, N., and Kimbell, R. (2022). A review of the valid methodological use of adaptive comparative judgement in technology education research. *Front. Educ.* 6. doi: 10.3389/feduc.2022.787926

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., and De Maeyer, S. (2017). Judging texts with rubrics and comparative judgement: taking into account reliability and time investment. *Pedagog. Stud.* 94, 283–303.

Heldsinger, S. A., and Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educ. Res.* 55, 219–235. doi: 10.1080/00131881.2013.825159

Heldsinger, S., and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *Aust. Educ. Res.* 37, 1–19. doi: 10.1007/BF03216919

Jones, I., and Alcock, L. (2012). "Summative peer assessment of undergraduate calculus using adaptive comparative judgement," in *Mapping University Mathematics Assessment Practices*, eds P. Iannone and A. Simpson (Norwich: University of East Anglia), 63–74.

Jones, I., and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educ. Stud. Math.* 89, 337–355. doi: 10.1007/s10649-015-9607-1

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6

Kimbell, R. (2021). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. *Int. J. Technol. Des. Educ*. doi: 10.1007/s10798-021-09654-w

Kimbell, R., Stables, K., Wheeler, A., Wosniak, A., and Kelly, V. (1991). *The Assessment of Performance in Design and Technology*. London: Schools Examinations and Assessment Council/Central Office of Information.

Lesterhuis, M., Van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., and De Maeyer, S. (2018). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Lit.* 18, 1–22. doi: 10.17239/L1ESLL-2018.18.01.02

Polanyi, M. (1958). *Personal Knowledge: Towards a Post–Critical Philosophy*. Chicago, IL: University of Chicago Press.

Pollitt, A. (2012a). Comparative judgement for assessment. *Int. J. Technol. Des. Educ.* 22, 157–170. doi: 10.1007/s10798-011-9189-x

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354

Seery, N., Canty, D., and Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *Int. J. Technol. Des. Educ.* 22, 205–226. doi: 10.1007/s10798-011-9194-0

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and de Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Princ. Policy Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542

Van Gasse, R., Lesterhuis, M., Verhavert, S., Bouwer, R., Vanhoof, J., Van Petegem, P., et al. (2019). Encouraging professional learning communities to increase the shared consensus in writing assessments: the added value of comparative judgement. *J. Prof. Cap. Commun.* 4, 269–285.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Princ. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Meas.* 42, 428–445. doi: 10.1177/0146621617748321

Whitehouse, C., and Pollitt, A. (2012). *Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment*. Manchester: Centre for Education Research and Policy.

Wiliam, D. (1998). "The validity of teachers' assessments," in *Paper Presented to Working Group 6 (Research on the Psychology of Mathematics Teacher Development) of the 22nd Annual Conference of the International Group for the Psychology of Mathematics Education*, Stellenbosch, 1–10.

Williams, P. J., and Kimbell, R. (2012). Special issue on e-scape [Special issue]. *Int. J. Technol. Des. Educ.* 22, 123–270.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.