



OPEN ACCESS

EDITED BY

Mohammed Salah,
A'Sharqiyah University, Oman

REVIEWED BY

Josefine Hofmann,
Gräbener Maschinentechnik GmbH &
Co. KG, Germany
Masibulele Phesa,
University of KwaZulu-Natal, South Africa
Ilona Rinne,
University of Gothenburg, Sweden

*CORRESPONDENCE

Kjetil Egelandstal
✉ kjetil.egelandstal@uib.no

RECEIVED 26 November 2025

REVISED 20 February 2026

ACCEPTED 23 February 2026

PUBLISHED 13 March 2026

CITATION

Egelandstal K and Færstad J-O (2026)
Beyond reliability: examining the
applicability of adaptive comparative
judgment in high-stakes assessment.
Front. Educ. 11:1755046.
doi: 10.3389/educ.2026.1755046

COPYRIGHT

© 2026 Egelandstal and Færstad. This is
an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Beyond reliability: examining the applicability of adaptive comparative judgment in high-stakes assessment

Kjetil Egelandstal^{1*} and Jan-Ove Færstad²

¹Department of Education & Centre for the Science of Learning and Technologies (SLATE), University of Bergen, Bergen, Norway, ²Faculty of Law, University of Bergen, Bergen, Norway

Introduction: Adaptive Comparative Judgment (ACJ) has been promoted as a promising alternative to criteria-based grading, yet evidence of its performance in high-stakes, text-intensive higher-education contexts remains limited. This study examines the applicability of ACJ as a grading method for complex, text-based assessments in higher education. Method: ACJ was applied *post hoc* to examination papers from two third-year law courses at a Norwegian university ($N = 107$) and compared with conventional grades using agreement statistics, discrepant-case expert review, examiner survey data, and judgment-time measures. Results: ACJ produced high reliability (Scale Separation Reliability = 0.85–0.89) and substantial grade convergence (> 90% identical or adjacent grades), although fewer than half of grades matched exactly and course-specific shifts in grade outcomes were sensitive to boundary placement. Discrepant-case reviews suggested that surface-level features (structure and readability) were often favored over deeper analytical reasoning, consistent with examiner reflections and short median judgment times.

Discussion: Translating rank orders into grades required substantial *post-hoc* thresholding and coordination work, and the absence of individualized justifications constrained transparency. Overall, while ACJ can yield reliable rankings, its practical use in high-stakes settings involving long written responses raises important validity and feasibility concerns.

KEYWORDS

adaptive comparative judgment, higher education, reliability, summative assessment, transparency, validity

Introduction

Grading complex, qualitative examinations presents persistent challenges in higher education, particularly in ensuring consistency, fairness, and defensibility across examiners. Even when rubrics, double marking, and moderation are employed, variation often arises from subjective interpretation and differing expectations (Sadler, 2009; Bloxham and Boyd, 2007). Such variation highlights the limits of purely criteria-based approaches and the need for assessment methods that can balance professional judgment, reliability, and transparency (Boud, 2007; Klenowski and Wyatt-Smith, 2014). In long, text-intensive examinations, these challenges are amplified by workload and time pressure, which can

constrain careful reading and make moderation procedures difficult to sustain consistently across large cohorts.

Adaptive Comparative Judgment (ACJ) has been proposed as an alternative to traditional criteria-referenced grading (Pollitt, 2012). Instead of assigning absolute grades, ACJ requires assessors to compare pairs of student responses and select the stronger; these judgments are aggregated algorithmically to produce a rank order of all submissions. As more judgments are made, the system targets comparisons between responses of similar quality, and each response is reviewed multiple times by different assessors.

Research on ACJ has consistently demonstrated high levels of reliability across contexts such as design, essay writing, and teacher education (Verhavert et al., 2019; Bartholomew and Jones, 2022; Jones and Wheadon, 2015). However, reliability alone does not ensure validity, transparency, or feasibility in high-stakes assessment contexts (Lesterhuis et al., 2022). While ACJ has mostly been applied in formative or experimental settings, its potential for summative use in complex, text-based disciplines such as legal education remains largely unexplored (Egelandsdal et al., 2025). Legal examinations require lengthy, multifaceted responses involving nuanced reasoning and structured argumentation, raising the question of whether ACJ can accommodate this complexity in a way that is both valid and practically manageable.

This study tested ACJ as a grading method in two third-year law courses - International Climate Law and Tax Law - at a Norwegian university. A total of 107 examination papers (59 in Climate Law and 48 in Tax Law, reflecting cohort paper availability) were reassessed using ACJ after initial grading through traditional methods. Nine examiners conducted pairwise comparisons of the examination papers using an online ACJ platform, after which grade boundaries were established *post-hoc* by course instructors based on the algorithmic rank order.

The study pursued two main aims: (1) to evaluate whether ACJ can be applied to assess complex, text-based examinations in a reliable and practically feasible manner within comparable resource constraints to traditional grading; and (2) to explore examiners' experiences of applying ACJ to such tasks. To address these aims, the study combines quantitative analysis of ACJ outcomes (reliability, agreement, and grade distributions) with qualitative evidence from discrepant-case reviews and examiner survey responses.

The guiding research question is:

How does Adaptive Comparative Judgment function as a grading method for complex, text-based examinations in legal education with regard to reliability, validity, transparency, and feasibility, and how do examiners experience its practical implementation?

By addressing these questions, the study contributes to ongoing discussions about alternative assessment models in higher education. First, it provides empirical evidence on ACJ's reliability and agreement with conventional grading in a high-stakes, text-intensive legal examination context. Second, it extends prior work by examining feasibility- and validity-relevant challenges that emerge when ACJ is implemented under realistic workload constraints, including judgment-time patterns, thresholding demands, and examiner experiences of transparency and accountability.

The remainder of the paper is structured as follows. The next section reviews prior research on ACJ with attention to reliability, validity, transparency, and feasibility in summative contexts. We then describe the study context, conventional grading baseline, ACJ design, and analytic approach. The results section presents reliability, agreement patterns, discrepant-case analyses, judgment-time findings, and calibration workload, followed by a discussion of implications for validity, fairness, transparency, and practical implementation in high-stakes assessment.

Previous research

In this study, we evaluate Adaptive Comparative Judgment (ACJ) through an assessment validity argument (Kane, 2013). Following an argument-based view of validity, evidence relevant to ACJ's use in high-stakes grading must extend beyond reliability to include construct relevance (what features judgments privilege), transparency and accountability (the defensibility of outcomes), and feasibility (the workload and standard-setting demands required for implementation). This perspective aligns with ACJ scholarship arguing that reliability alone is insufficient for summative use (Kelly et al., 2022; Lesterhuis et al., 2022). It guides our review below and motivates our focus on reliability, agreement with conventional grades, discrepant-case patterns, examiner experiences, and the practical work involved in translating rank orders into grades.

Reliability

Adaptive Comparative Judgment (ACJ), developed from Thurstone (1927) psychometric law of comparative judgment, has evolved from early applications in language proficiency testing (Pollitt and Murray, 1993) to the assessment of performances across educational domains (Hartell and Buckley, 2021). Across these studies, ACJ has consistently demonstrated high internal reliability, with Scale Separation Reliability (SSR) values typically exceeding 0.80 and, in some cases, reaching 0.95 (Verhavert et al., 2019; Bartholomew and Jones, 2022). SSR expresses how consistently assessors distinguish stronger from weaker performances. However, high reliability may conceal examiner disagreement, since adaptive algorithms can over-optimize pairings and thus inflate consistency estimates (Bramley and Vitello, 2018).

Achieving such reliability generally requires a substantial number of comparisons. Meta-analyses indicate that 10–14 comparisons per script are needed to exceed SSR 0.70, and 25–37 to reach 0.90 (Verhavert et al., 2019). For example, McMahon and Jones (2014) found that five science teachers required 14 h to reach SSR 0.87 when using comparative judgment to evaluate 154 short responses, compared to just 3 h using traditional marking. Even when accounting for double-marking, which would have required approximately 6 h, the comparative judgment process still took more than twice as long. Consistent with these findings, a recent study within higher education (Egelandsdal et al., 2025) showed that the number of comparisons completed fell short of the threshold required for a reliable ranking. This was likely due not

to examiners' capacity, but to the assumption that their workload should mirror traditional grading expectations. Clearer instructions and firmer guidance may support better reliability in practice. These findings highlight that while ACJ is statistically promising, its feasibility hinges on institutional expectations and the structure of examiner engagement.

Validity and grading

While ACJ provides reliable rank orders, its validity rests on the extent to which comparative decisions reflect intended learning outcomes. Unlike criteria-based grading, ACJ relies on holistic, relative comparisons rather than predefined rubrics or analytic criteria. Research shows that assessors draw on internalized conceptions of quality, and that combining perspectives can capture the multidimensional nature of complex constructs (Lesterhuis et al., 2022; Buckley et al., 2020). Yet emphasis differs: some assessors privilege structure and linguistic fluency, while others prioritize reasoning and argumentation (Lesterhuis et al., 2022; van Daal et al., 2019). Less experienced examiners tend to rely more on surface features, whereas experienced assessors focus on deeper conceptual aspects (Egelandsdal et al., 2025).

A major limitation of ACJ in summative contexts is that it produces relative rank orders rather than discrete grades. Translating ranks into grades often entails norm-referenced assumptions, which can disadvantage cohorts performing above or below the mean. To mitigate this, thresholding procedures have been proposed, using anchor or boundary scripts to convert continuous rankings into grade bands (Steedle and Ferrara, 2016; Marshall et al., 2020). Yet in many university settings, including professional fields such as law, examination questions change each semester, making the reuse of fixed anchor scripts impractical.

Accountability and transparency also remain problematic. Because ACJ outcomes are derived relationally, they provide limited audit trails for explaining how criteria are applied, complicating grade justification and defense (Kelly et al., 2022). In many higher-education systems, assessment decisions must be justified through explicit criteria and individual feedback. ACJ instead generates aggregated rankings without individualized rationales, raising concerns about fairness and defensibility (Egelandsdal et al., 2025).

Examiners' perspectives and feasibility

Such concerns are not purely procedural but affect examiners' experiences and perceptions of professional responsibility. The absence of explicit evaluative criteria and individual justifications may limit external transparency while also altering examiners' sense of accountability for their decisions. In high-stakes contexts, this shift from individual to collective grading logic can affect both engagement and judgment quality. Understanding examiner perspectives is therefore crucial to evaluating the practical feasibility of ACJ in real assessment systems.

Prior studies report that examiners often appreciate the intuitive, holistic nature of pairwise comparisons and the perceived

authenticity of ACJ compared with rigid rubrics (van Daal et al., 2019; Verhavert et al., 2019). Yet the large number of required judgments can cause cognitive fatigue (Steedle and Ferrara, 2016; Jones and Davies, 2024), especially in disciplines where responses are long and analytically dense. Under such pressure, assessors may begin to prioritize efficiency over depth, potentially leading to more superficial decisions (Egelandsdal et al., 2025).

Another recurring issue concerns how examiners interpret their roles in a system that yields group-level rankings without individual evaluative statements. While traditional approaches enable assessors to justify each grade with reference to explicit criteria, ACJ offers only a relative position in an ordered set (Kelly et al., 2022). Examiners in professional fields such as law have expressed discomfort with this lack of individual accountability, noting that it constrains feedback and post-assessment dialogue (Egelandsdal et al., 2025). Such limitations may also reduce examiners' own sense of ownership over grading decisions, potentially influencing judgment consistency.

Collectively, prior research indicates that the success of ACJ depends not only on psychometric robustness but also on its perceived fairness, transparency, and workload implications for examiners. The present study therefore evaluates these dimensions empirically through both performance data and examiner reflections from two university law courses.

Methods

Context and participants

This study was conducted at a Norwegian university using examination papers from two third-year law courses: Tax Law (48 papers) and International Climate Law (59 papers). The number of papers reflects the number of eligible examination papers available in each course cohort. We did not sub-sample to equalize group sizes, as the aim was to evaluate ACJ under realistic implementation conditions. The papers were assessed by nine experienced examiners (four in Tax Law and five in Climate Law). All examiners were legal professionals familiar with course content and grading practices. Each course had a course coordinator (two coordinators in total) responsible for the examiner guide and oversight of the conventional grading process; these coordinators also conducted the *post-hoc* grade calibration of the ACJ rank orders (see "Grade calibration"). In addition, one independent expert examiner contributed to a structured discrepant-case review together with the course coordinator of each course and an extended review of one-grade discrepancies alone; this expert did not participate in the original grading or the ACJ judging.

The examination papers were included if they were part of the relevant written examination component and were eligible for research use under the course information/consent procedure (see Ethics). Examiners were recruited from the ordinary pool of course examiners used in conventional grading; participation in the ACJ study was voluntary and based on informed consent.

Each paper had first been assessed through traditional criterion-referenced grading. In this process, each paper was graded holistically by a single examiner with reference to an examiner

guide, which provided qualitative descriptors of performance levels (A–F), relevant learning outcomes and course literature, and question-specific guidance on expected approaches and common pitfalls. Double marking was not used as a routine procedure; instead, moderation focused on borderline and/or weak papers. For quality assurance, the course instructor conducted level control by comparing grades across examiners and reviewing papers in the weakest range of each passing grade (and any paper flagged as potentially unsatisfactory). Papers considered at risk of failure were subject to secondary review by the course coordinator, and a failing grade was assigned only after agreement was reached between the examiner and coordinator. In the conventional process, grade boundaries are normally set through these qualitative performance descriptors and local standard-setting practices rather than fixed numerical cut scores, with distribution checks and targeted moderation used to ensure that grade levels are applied consistently across examiners. Accordingly, the number of papers reviewed in moderation may vary across cohorts and examiners. The results from the traditional grading served as the baseline for comparison with ACJ results.

Tax Law (10 ECTS) was assessed through a 6-h written examination consisting of three problem scenarios; students answered four questions, typically producing 3,500–4,000 words. International Climate Law (10 ECTS) combined a take-home assignment (40%) and a four-hour written examination (60%), but only the written examination was included here. Responses ranged from about 1,000 to 3,000 words. Both assessments were summative course examinations contributing to students' final course results; however, the ACJ exercise was conducted *post hoc* and had no impact on official grades. Examiner guides outlining learning outcomes and expected approaches were available in both courses. The two courses were selected as they both are third-year courses where examinations are relatively long and analytically demanding, providing a stringent test of ACJ's applicability to complex, text-based summative assessment. Examinations in these courses are, however, not as comprehensive as in fourth- or fifth-year courses.

Only one of the nine examiners had prior ACJ experience. All received a written briefing and access to examiner guides, though they were not required to follow them strictly during the ACJ task, consistent with ACJ's emphasis on holistic judgment.

ACJ assessment design

We used the *RM Compare* platform (RM, 2024) to implement the Adaptive Comparative Judgment method. *RM Compare* is a web-based comparative judgment platform that presents papers in pairs, records assessors' decisions (and optional comments), and uses an adaptive pairing algorithm to refine an estimated rank order as judgments accumulate. It also outputs standard ACJ statistics (e.g., Scale Separation Reliability) and logs process data such as decision times. Examiners reviewed pseudonymized student responses in pairs and selected the stronger paper. The algorithm adaptively refined the rank order by pairing papers of similar estimated quality as more judgments accumulated.

Each examiner completed 90 pairwise comparisons, yielding a total of 360 judgments across four examiners in Tax Law and 450

judgments across five examiners in Climate Law. This number was determined using *RM Compare*'s judgment calculator to achieve at least 15–16 comparisons per paper - sufficient for Scale Separation Reliability (SSR) > 0.8 (Verhavert et al., 2019). Judging was concluded when this pre-specified allocation was completed (i.e., when all examiners had completed 90 comparisons each).

In the Tax Law course, examiners provided brief justifications for each decision, whereas no justifications were required in Climate Law. This enabled a comparison of judgment behavior and examiner experiences across the two implementations.

Decision times were logged to allow analysis of examiner pace. *RM Compare* records a decision-time value for each pairwise judgment, defined here as the elapsed time from when a comparison pair is opened/presented to the examiner until the examiner submits the A/B decision (and, where enabled, submits a written justification). Because justifications in Tax Law were entered in the same interface prior to submission, the logged time includes time spent writing them. *RM Compare* reports both mean and median decision times; because the time logs do not distinguish active working time from idle time (e.g., a comparison left open during a break), we report the median decision time per examiner rather than the mean.

Since the course contexts also differed in examination format and response length, we report these contrasts as contextual differences rather than attributing them to justification requirements.

Grade calibration

After the ACJ rankings were finalized, course coordinators set grade boundaries by locating threshold positions along the rank-ordered list, guided by the distribution of ACJ parameter values. Boundary setting was conducted blind to the conventional grades. The coordinators inspected the parameter plot for visible discontinuities ("elbows") and treated each plausible inflection point as a candidate cut score - particularly for the B/C and C/D boundaries, where more than one elbow was sometimes defensible. To validate candidate cut points, coordinators then conducted targeted qualitative reviews of papers immediately around each boundary, iterating between plot inspection and close reading until the boundaries were judged to reflect plausible grade-level transitions in light of the examiner guide and local performance standards.

In Tax Law, this review focused on (i) the A/B transition at a clear elbow (between the two adjacent papers at that point), (ii) three papers in the B/C borderland, and (iii) five papers spanning two candidate C/D elbows; the D/E transition was explored but remained ambiguous. In Climate Law, where the plot showed few clear elbows, coordinators compared multiple papers above and below each tentative boundary (including around two candidate A/B cut points), checked for local anomalies where subsequent papers appeared stronger than the first paper "below" a cut, and relied more heavily on close reading; the lowest-end D/E/F boundaries were supported by a steeper decline in the plot.

This procedure resembles a retrospective "boundary script" approach (Steedle and Ferrara, 2016) without predefined

exemplars. After boundaries were finalized, ACJ-derived grades were assigned and compared with original grades to examine agreement and discrepancies. Finally, we conducted a simple sensitivity check by shifting each boundary by ± 1 -2 rank positions and recalculating net grade shifts (Table 4).

Structured review of two-grade discrepancies

To explore discrepancies between ACJ and traditional grading, we focused on examination papers with a two-grade or greater difference between the two methods (four in Tax Law and two in Climate Law, since only these papers met the ≥ 2 -grade criterion; we therefore reviewed all eligible cases).

These were subjected to a structured expert review conducted by what we refer to as the Expert Review Team. Each team comprised two members: the course coordinator and an independent expert examiner who had not participated in the original grading but had substantial experience assessing law examination papers. The same independent expert examiner participated in both teams. Thus, two course-specific teams were formed - one for Tax Law and one for Climate Law.

- *Independent Evaluation*: The coordinator and the examiner independently assessed the papers using conventional grading methods, without consulting either the original or ACJ-assigned grades. They graded the papers holistically based on their understanding of performance standards.
- *Collaborative Discussion*: The reviewers then compared their judgments, discussed strengths and weaknesses of each paper, and arrived at a shared evaluation.

This qualitative review helped identify assessment features that may have contributed to over- or under-ranking by ACJ.

Extended review of one-grade discrepancies

In addition to the structured review of two-grade discrepancies, the same experienced examiner conducted a broader analysis of all papers that differed by one grade between ACJ and traditional grading. Rather than regrading each paper, this review identified general patterns by examining textual and structural features (e.g., clarity, argumentation, organization, and placement of key points) that appeared to recur in papers receiving higher or lower grades under ACJ. The goal was to explore whether such characteristics might help explain systematic tendencies in the observed discrepancies.

Data collection and analysis

Data were collected and analyzed from four complementary sources:

- *ACJ Process Data*: The RM Compare platform logged examiner judgments, decision times, and agreement metrics. These data enabled analysis of reliability (SSR), examiner pace, and convergence with the consensus ranking.
- *Coordinator reports on Grade Calibration*: Written reports from course coordinators documented their interpretation of inflection points in the ranking graphs, how threshold papers were selected, and the challenges of reconciling ACJ rankings with existing grading standards. These reflections offered insight into the interpretive work involved in *post-hoc* grade mapping and highlighted both the potential and ambiguity of visual cues in the ACJ output.
- *Discrepant Case Review*: The qualitative expert review of outlier cases helped interpret where and why the ACJ ranking diverged from traditional assessment.
- *Examiner Survey*: After completing ACJ, all examiners completed an anonymous open-ended survey via *SurveyXact* (Xact by Ramboll, 2024). The survey was distributed the day after completing the ACJ and all examiners responded within a week. The survey comprised one brief background question about prior examiner experience at the faculty and within the relevant course area, followed by seven open-ended prompts asking examiners to describe: (1) their overall experience of comparative judging (what worked well and what was challenging), (2) how pairwise judging compared to evaluating one script at a time, (3) whether comparative judging led them to emphasize different aspects of student performance than in conventional criteria-based grading (and why), (4) whether comparative judging affected how differentiated their judgments were relative to conventional grading, (5) how they experienced completing 90 pairwise judgments and whether the time available was sufficient for quality judgments, and (6) their estimated time per judgment and total time spent, plus (7) an optional final “anything else” prompt. Responses were provided as free-text without word limits.

Survey data were analyzed using reflexive thematic analysis (Braun and Clarke, 2006). The analytic procedure involved: (1) familiarization through repeated reading; (2) prompt-informed inductive coding of the full dataset; (3) clustering codes into themes; (4) iterative theme refinement by checking themes against the coded extracts and the full dataset; and (5) defining and naming themes for reporting. Although coding was inductive, the open-ended prompts necessarily shaped the topical scope of responses; codes and themes were nevertheless developed from the content and meaning of the responses rather than from a pre-specified coding scheme. Coding was conducted by the main author, with the second author reviewing the coding structure and theme definitions; disagreements were resolved through discussion and refinement of the themes.

Ethics

Formal ethics approval was not required for this study under Norwegian regulations for educational research using pseudonymized examination materials and anonymous staff survey

data and no impact on students' grades. The project was therefore handled through the University of Bergen's institutional data-protection procedures and internal research project registration system (Rette). Students were informed that their anonymized examination responses could be used for research and were given the opportunity to opt out. The ACJ process took place after final grades were released and had no effect on students' results.

All student papers were pseudonymized with numeric identifiers. Examiners saw no personal data beside the pseudonymized student papers. Both researchers and examiners were unable to re-identify the students. The study was carried out under GDPR Article 89(1), allowing pseudonymized research without further legal basis, as per Article 5(1)(b).

Examiners gave informed consent and were fully aware that their judgments and survey responses would be anonymized and analyzed for research purposes. Survey participation was anonymous, and consent was provided through voluntary participation in the survey.

Results

Coordinator reflections on grade boundaries

Course coordinators in both Tax Law and Climate Law reported that converting ACJ rank orders into grade boundaries required substantial *post-hoc* interpretive work. In Tax Law, a relatively clear boundary was identified between A and B, other transitions (e.g., B/C and C/D) were more ambiguous, requiring extensive cross-reading of neighboring responses. The coordinator noted that "very few points stood out as decisive," and the process ultimately resembled standard grading in its level of effort - taking nearly 10 h to complete.

Across both courses, coordinators described grade mapping as an interpretive step that required combining the ACJ parameter plot with targeted re-reading of papers around candidate boundaries. In Tax Law, the ranking graph provided clearer discontinuities, making tentative boundary positions easier to identify. In Climate Law, the graph showed a more gradual decline, so avoiding misclassification required iterative checks of multiple papers above and below each tentative cut-off. Overall, the reflections suggest that ACJ rankings can provide a useful starting point for *post-hoc* standard setting, but they do not remove the need for close reading and expert judgment - especially when the parameter plot lacks clear 'elbows'.

Comparison of ACJ outcomes with traditional grading

The ACJ method produced fully ordered rankings of examination papers in both Climate Law and Tax Law. Reliability was high, with Scale Separation Reliability (SSR) scores of 0.89 and 0.85, respectively, consistent with established thresholds in prior ACJ research (e.g., Verhavert et al., 2019).

ACJ-derived grades showed moderate to substantial agreement with the original grades assigned through traditional assessment. Tables 1, 2 report confusion matrices (original grade × ACJ grade) for each course, and Table 3 reports grade distributions under both methods. Overall agreement was moderate to substantial. In Climate Law ($N = 59$), exact agreement was 49.2% and agreement within ± 1 grade was 96.6%; weighted κ was 0.61 (linear) and 0.79 (quadratic), and ACJ grades were strongly associated with original grades (Spearman $\rho = 0.79$, $p < 0.001$). In Tax Law ($N = 48$), exact agreement was 41.7% and agreement within ± 1 grade was 91.7%; weighted κ was 0.44 (linear) and 0.66 (quadratic), with a moderate-to-strong association between ACJ and original grades (Spearman $\rho = 0.64$, $p < 0.001$).

The confusion matrices further show that most discrepancies were small in magnitude and concentrated around adjacent grade categories (Tables 1, 2). Consistent with this pattern, more than nine in ten papers in both courses received an ACJ grade that was either identical to, or within one grade level of, the original grade.

Grade distributions showed small but directionally different shifts under ACJ (Table 3). In Tax Law, 22.9% of papers received a higher grade under ACJ, compared to 35.4% receiving a lower grade (net -12.5 percentage points). In Climate Law, 32.2% received a higher grade under ACJ compared to 18.6% receiving a lower grade (net $+13.6$ percentage points). However, these net shifts were sensitive to grade-boundary placement: shifting each boundary by $\pm 1-2$ rank positions produced noticeable changes in the estimated net shift (Table 4). Taken together, the results indicate substantial convergence between ACJ and conventional grading, while suggesting that any directional "shift" depends partly on boundary choices.

Discrepancies and patterns in judgment

Qualitative analysis of the most discrepant cases (two-grade differences) showed a tendency for ACJ to favor papers with strong structure, clarity, and presentation - even when the substantive analysis was weaker. This pattern was particularly evident in the four Tax Law cases with significant grading discrepancies. One candidate received an A in ACJ but a C in traditional grading. While the Expert Review Team noted a very clear, almost schematic, structure, they described the legal reasoning as superficial and inconsistent. The strongest parts of the paper were found at the beginning, with quality declining in later sections. The coordinators suggested that the paper gave the impression of content having been memorized rather than understood, and that the accessible structure may have boosted its ACJ ranking despite its uneven content. In their independent reviews, one expert reviewer placed the examination paper at a strong D, while the other arrived at a weak C.

In contrast, three candidates were downgraded in ACJ compared to their original grades. Two of these had received A and C in traditional assessment but were ranked as C and E, respectively, through ACJ (the expert review team judged the first paper as B- and B/A, and the second as C- and D). In both cases, the Expert Review Team noted evidence of relevant legal reasoning and familiarity with key concepts, but also that

TABLE 1 Confusion matrix for Climate Law (original grade × ACJ grade).

Original grade	A	B	C	D	E	F	Total
A	6	3	0	0	0	0	9
B	7	3	3	0	0	0	13
C	1	4	11	3	0	0	19
D	1	0	4	6	1	0	12
E	0	0	0	2	2	1	5
F	0	0	0	0	0	1	1
Total	15	10	18	11	3	2	59

N = 59. Rows indicate original grades; columns indicate ACJ-derived grades.

TABLE 2 Confusion matrix for Tax Law (original grade × ACJ grade).

Original grade	A	B	C	D	E	F	Total
A	2	1	1	0	0	0	4
B	3	3	7	0	0	0	13
C	1	2	10	4	2	0	19
D	0	0	5	3	2	0	10
E	0	0	0	0	1	0	1
F	0	0	0	0	0	1	1
Total	6	6	23	7	5	1	48

N = 48. Rows indicate original grades; columns indicate ACJ-derived grades.

the substantive quality was uneven and at times underdeveloped; moreover, their presentations were dense, indirect, or overly narrative. They further emphasized that one of the papers was clearly stronger overall, whereas the other combined this dense style with more pronounced weaknesses in both substance and form. This contrast illustrates how within-band variation in C/D papers may interact with presentation features in fast-paced comparative judgments, making key analytical points harder to identify without close reading, potentially disadvantaging such responses in fast-paced comparative judgments. The third downgraded candidate was marked C in traditional grading but ranked E through ACJ. While the Expert Review Team acknowledged some basic subject knowledge, they pointed to significant misunderstandings, vague formulation of the legal issues, and a disjointed structure. They noted that the paper's strengths were easy to overlook due to its poor structure, while its weaknesses were immediately apparent. They also noted that the paper was significantly shorter in length compared to the other papers assessed. In their independent reviews, one expert reviewer placed the examination paper at a D, while the other arrived at a C.

A similar pattern appeared in two discrepant Climate Law cases: both of which received a grade of A through ACJ, despite having been assigned C and D in traditional grading. The Expert Review Team noted that both papers, while well-structured, lacked analytical depth and rigor in legal reasoning. Although the quality of both papers fluctuated throughout the text, both answered the first question relatively well. In the first case, the expert reviewers

TABLE 3 Grade distributions under traditional grading and ACJ.

Course	Grade	Original <i>n</i> (%)	ACJ <i>n</i> (%)	Total <i>N</i>
Climate Law	A	9 (15.3)	15 (25.4)	59
	B	13 (22.0)	10 (16.9)	
	C	19 (32.2)	18 (30.5)	
	D	12 (20.3)	11 (18.6)	
	E	5 (8.5)	3 (5.1)	
	F	1 (1.7)	2 (3.4)	
Tax Law	A	4 (8.3)	6 (12.5)	48
	B	13 (27.1)	6 (12.5)	
	C	19 (39.6)	23 (47.9)	
	D	10 (20.8)	7 (14.6)	
	E	1 (2.1)	5 (10.4)	
	F	1 (2.1)	1 (2.1)	

Percentages are column percentages within each course.

placed the examination paper at D and weak C, respectively. In the second case, both reviewers placed the paper at C.

These findings were consistent with findings from the review of the one-grade discrepancies. The reviewers found that half of the papers that performed better under ACJ were characterized by a fluent, readable style. Of the 26 papers across both courses that received a higher grade in ACJ compared to conventional grading, 13 were evaluated positively on this marker, 9 received a neutral score, and only 4 were evaluated negatively. Conversely, the reviewer identified an even stronger tendency among the papers that performed worse under ACJ than under traditional grading. Most of these were characterized by a lack of fluency and readability. As many as 17 of the 25 papers in this category received a negative score on this marker, while the remaining 8 were scored neutral. A similar pattern was observed regarding structure. Of the 26 papers that performed one grade better in ACJ, 16 scored positively on accessible structure. In contrast, of the 25 papers that performed one grade worse in ACJ, 15 scored negatively on this marker.

Taken together, these divergent examples underscore a recurring vulnerability in ACJ: that surface-level accessibility and presentation can unduly influence comparative assessments when judgments are made quickly across a high volume of comparisons, sometimes at the expense of deeper disciplinary understanding.

Judgment time

Analysis of examiner behavior revealed substantial variation in judgment time across and within the two ACJ implementations. In the Climate Law ACJ, where no written justification was required, decisions were generally made rapidly. For four out of five examiners, the median time per comparison was approximately 1.5 min or less; two completed judgments in as little as 18–24 s on average. One examiner was a clear outlier, with a median time around 12 min. Overall, aside from this individual, Climate Law

TABLE 4 Sensitivity of net grade shift to boundary placement (\pm rank positions per boundary).

Course	-2 ranks	-1 rank	Baseline	+1 rank	+2 ranks
Tax Law (N = 48)	-25.0	-18.8	-12.5	-2.1	+6.3
Climate Law (N = 59)	-1.7	+6.8	+13.6	+16.9	+25.4

Values are net shift in percentage points (% upgraded - % downgraded) after shifting each grade boundary by the specified number of rank positions.

judgments were typically completed in under 90 s, and in some cases under 30 s.

In the Tax Law ACJ, where a brief written justification was required for each judgment, decision times were markedly longer. The four Tax Law examiners had median comparison times of approximately 2 min, 8.5 min, 12 min, and 22 min, respectively. Even the fastest examiner in Tax Law was slower than the majority of Climate Law examiners.

Tax Law responses were also substantially longer - around 3,500–4,000 words on average, compared to 1,000–3,000 words in Climate Law - which may also have contributed to the extended reading times. However, these between-course differences in justification requirements and response length should be interpreted descriptively rather than causally, as the two contexts may have differed in several uncontrolled factors.

From a workload perspective, these time differences are notable. Assuming 90 comparisons per examiner, the fastest Climate Law examiners likely spent around 2 to 4.5 h in total, whereas the slower Tax Law examiners may have spent 10 to 15 h or more. This remained below the 35 h each examiner was compensated for, based on traditional grading expectations.

Examiners' reflections on ACJ

Survey responses revealed five key themes:

1. *Efficiency and Practicality*: Examiners found pairwise judgments easier than holistic grading but noted the cumulative burden of many decisions. Several described the task as “mechanical” or “mentally tiring.”
2. *Difficulty of Direct Comparison*: Examiners expressed discomfort with binary decisions in cases where two papers had different but comparable strengths.
3. *Shift in Assessment Focus*: Many admitted that presentation and structure influenced decisions more than intended. ACJ focus on quick holistic impression, may favor readability over deeper understanding.
4. *Time Pressure and Fatigue*: Some examiners admitted to partial readings, particularly in the Climate Law implementation (where no justifications were required). This raised concerns about fairness and attention to full content.
5. *Fairness and Future Use*: Examiners were divided on ACJ's suitability for high-stakes use. While some appreciated the reliability due to collective judgments of several examiners, others highlighted the lack of transparency and individual feedback as problematic. Most supported using ACJ as a complementary tool rather than a full replacement, suggesting

it could enhance formative assessment or support calibration and moderation processes.

Discussion

While ACJ produced reliable rankings and broad convergence with traditional grades, closer analysis revealed tensions between consistency, validity, and practical implementation. Key concerns included a potential overemphasis on surface-level features, fast judgment speed, limited transparency, and difficulties translating rankings into defensible grades. Additionally, some examiners saw potential in using ACJ for purposes such as calibration and moderation, though these ideas emerged as forward-looking reflections rather than tested applications. In line with an argument-based view of validity (Kane, 2013), we interpret these results as evidence bearing on multiple aspects of defensible summative use - not only reliability, but also construct relevance, transparency/accountability, and feasibility. The following sections examine these issues in depth, focusing on the method's performance across dimensions of reliability, validity, judgment processes, fairness, and feasibility.

Reliability vs. construct validity

ACJ yielded strong reliability scores ($SSR = 0.85-0.89$), reinforcing previous research that demonstrates its capacity to differentiate quality consistently (Verhavert et al., 2019; Bartholomew and Jones, 2022). While over 90% of grades were either identical to or within one grade of traditional assessments, fewer than half matched exactly, and agreement was moderate to substantial when assessed using weighted kappa and Spearman's ρ . High ACJ reliability alongside moderate exact grade agreement is consistent with prior studies showing that stable rank ordering can still yield uncertainty when ranks are translated into categorical grades (e.g., Pollitt, 2012; Bramley and Vitello, 2018; Kelly et al., 2022).

Because conventional grades are a reference point rather than a gold standard, convergence is interpreted as one strand of validity evidence; the key question is what ACJ tends to advantage when outcomes diverge.

To understand what may drive these differences, we examined divergent cases and examiner reflections. This qualitative material suggests that the high reliability observed may reflect shared attention to surface-level features - particularly clarity, organization, and readability - rather than deeper legal reasoning. This pattern raises construct-validity concerns in contexts where analytical depth is central. Taken together, the discrepancies

suggest that examiners, under pressure to make many holistic comparisons, often favored accessible presentation over substantive reasoning. Survey responses echoed this tendency, with examiners acknowledging the risk of overvaluing surface clarity at the expense of legal depth. Concerns that comparative judgments may overweight presentation-related cues relative to disciplinary substance have also been noted in prior work, especially where tasks are complex and responses are long, reinforcing that construct relevance must be examined alongside reliability in summative uses (Chambers and Cunningham, 2022; Kelly et al., 2022).

We also observed small but directionally different shifts in ACJ-derived grades across the two courses (Table 3), and these shifts were sensitive to *post-hoc* boundary setting (Table 4). In Tax Law, the baseline mapping yielded more downgrades than upgrades, whereas Climate Law showed the opposite pattern. However, the sensitivity check demonstrated that shifting boundaries by only ± 1 – 2 rank positions can meaningfully change the estimated net shift. This indicates that any directional “shift” should be interpreted cautiously and underscores that observed differences between ACJ-derived and conventional grades reflect both (i) how judgments are expressed in pairwise comparisons and (ii) how ranks are translated into grade categories through local standard-setting.

This underscores the risk that ACJ, in its current form, privileges presentation over construct-relevant disciplinary competence (Chambers and Cunningham, 2022; Egelandsdal et al., 2025). Addressing these challenges may require calibration mechanisms, targeted examiner training, or structured justifications to ensure that core competencies - such as legal reasoning - are adequately weighted in the judgment process. Yet any such measures are likely to increase time, effort, and cost, potentially offsetting ACJ's efficiency advantage in large-scale assessment.

Judgment speed and cognitive demands

Our analysis revealed marked variation in how long examiners spent on ACJ decisions, underscoring a fundamental tension between efficiency and depth. A key result is that a large share of judgments were made very quickly. In Climate Law, decisions were often rapid: for four out of five examiners, the median time per comparison was approximately 1.5 min or less, and two examiners completed judgments in as little as 18–24 s on average. Aside from one outlier, Climate Law judgments were typically completed in under 90 s, and in some cases under 30 s. In Tax Law, judgment times were generally longer and more variable across examiners.

These time patterns raise concerns about depth and fairness: multi-page legal examination papers cannot realistically be assessed thoroughly in under a minute, which increases reliance on surface cues (structure, tone, fluency). In such rapid holistic comparisons, primacy and halo effects may become more influential (Steiner and Rain, 1989; Murphy et al., 1993), potentially advantaging papers that signal clarity early while disadvantaging responses whose strengths require sustained reading.

Consistent with prior work, fast judgments do not necessarily undermine accuracy on average (Pollitt, 2012; Jones and Wheadon,

2015). In our data, time use showed no clear relationship with reliability or grade convergence across courses; however, because the two implementations differed in response length and format, these contrasts are interpreted descriptively rather than causally. Importantly, the discrepant-case reviews still indicate that rapid comparisons can disadvantage papers whose strengths are less immediately visible.

Finally, contextual factors must be acknowledged. Examiners knew that ACJ was part of a research project and would not affect actual student grades. This “shadow assessment” context may have influenced behavior in either direction - prompting some to make quicker decisions and others to take extra care. As this could not be controlled, uncertainty remains about how judgment times might translate to authentic high-stakes settings.

Converting ACJ rankings into grades

While ACJ produces a reliable rank order of student responses, converting this into defensible grades remains a key challenge. Coordinators in both courses found only the A/B threshold relatively clear; other boundaries required extensive manual reading of neighboring responses, closely resembling traditional grading practices. This reflects a structural limitation of ACJ: the method yields relative performance data but not criterion-referenced grade levels. Without predefined benchmarks, thresholds had to be established through local professional judgment, placing a heavy workload on course coordinators. Thus, although ACJ produces a reliable rank order, the final step of assigning grades still demands substantial interpretive effort, particularly when performance deviates from expectations. This dependence on local thresholding is also reflected in the sensitivity check (Table 4), which shows that small boundary shifts can meaningfully alter the net pattern of grade changes.

Transparency, feedback, and fairness

One of the most significant limitations of ACJ is its lack of transparent, individualized feedback. In many higher education contexts, students are entitled to explanations for their grades, particularly in high-stakes assessments. ACJ, however, only generates relative rankings, offering no direct rationale for individual outcomes. This makes it difficult to justify grades or provide meaningful feedback - an issue that was repeatedly raised by examiners in our study.

Some workarounds are possible. For example, a coordinating professor could review the ACJ ranking and provide feedback on individual papers, drawing on judge justifications recorded in the system. However, this approach reintroduces much of the manual workload that ACJ is intended to reduce, thereby undermining its main practical advantage. Furthermore, although this may provide students with meaningful feedback, it does not constitute a justification of the grade awarded, as it is formulated somewhat independently of the grading process.

Another possibility is the use of AI to provide automated feedback, either directly from student papers or in combination

with ACJ audit trails. While such tools can generate text-level comments, they do not resolve the core problem: ACJ grades remain based on relative rankings rather than explicit criteria. As a result, students may still lack a clear justification for their grade, echoing broader critiques that ACJ without meaningful feedback mechanisms risks falling short of the expectations of summative assessment (Kelly et al., 2022).

Given these limitations, institutions adopting ACJ must consider how to supplement it with mechanisms that ensure students receive clear, criterion-referenced explanations of their performance - especially when used in contexts where grades carry significant consequences.

ACJ as a complementary assessment tool

While examiners in this study expressed clear reservations about using Adaptive Comparative Judgment (ACJ) as a primary method for grading high-stakes legal examinations, many nevertheless identified potentially valuable complementary applications - particularly for examiner calibration and *post-hoc* moderation. Several examiners envisaged using ACJ to establish shared performance standards before marking begins, especially in settings with multiple markers. Others suggested that ACJ could function as a structured way of flagging borderline or contested cases for closer review after initial grading. These reflections are consistent with prior work highlighting ACJ's capacity to surface differences in judgment and support examiner calibration and cross-assessor consistency (Pollitt, 2012; Bramley and Vitello, 2018). Our contribution is to show how these potential benefits are viewed by examiners in a high-stakes legal context, while also illustrating the practical conditions (thresholding workload and transparency constraints) that may shape whether such uses are feasible in practice. However, it is important to note that our study did not systematically evaluate these specific uses, and we therefore treat them as plausible applications rather than evidence-based recommendations. If ACJ is to be implemented in this way, careful consideration is needed to ensure that the additional workload and coordination demands do not outweigh the intended benefits. Finally, it is possible that emerging AI-based tools may offer alternative supports for calibration or case flagging, but such approaches raise distinct concerns related to bias, transparency, and explainability (van den Berg and Papadopoulos, 2024).

Conclusion

The aim of this study was to evaluate whether Adaptive Comparative Judgment (ACJ) can be applied to assess complex, text-based law examinations in a reliable and practically feasible manner under realistic resource constraints, and to examine examiners' experiences of its use in a high-stakes summative assessment setting. Across two third-year law courses, ACJ generated reliable rankings (SSR = 0.85–0.89) and broad convergence with conventional grading: over 90% of papers received either the same grade or differed by only one grade,

although fewer than half matched exactly. At the same time, course-specific net shifts in grade outcomes were sensitive to grade-boundary placement, underscoring that any directional differences between ACJ-derived and conventional grades should be interpreted cautiously and as partly contingent on local standard-setting choices.

The study's main contribution is to show that strong reliability can co-exist with practical and validity-relevant challenges when ACJ is used for long, analytically demanding, high-stakes examination papers. Discrepant-case reviews indicated that surface-level features (clarity, organization, readability) were often rewarded over deeper legal reasoning, raising concerns about construct relevance in contexts where analytical depth is central. Feasibility and defensibility also emerged as key constraints: translating rank orders into criterion-referenced grades required substantial *post-hoc* thresholding and coordinator effort, and the limited availability of individualized rationales constrained transparency and the ability to defend outcomes in appeals.

Several limitations should be noted. The study was conducted in two courses within one institutional setting, and the ACJ exercise was *post hoc* ("shadow assessment"), which may have influenced examiner behavior. The two course contexts also differed in examination format and response length, limiting causal interpretation of between-course contrasts. In addition, grade calibration relied on local professional judgment; although the sensitivity check strengthens transparency around boundary dependence, alternative standard-setting approaches may yield different aggregate patterns.

On this basis, we do not recommend ACJ in its current form as a stand-alone grading method for high-stakes, text-intensive examinations.

For practice, institutions considering ACJ in such settings must anticipate substantial thresholding workload and develop complementary mechanisms to ensure transparency, defensibility, and criterion-referenced explanation of grades.

For research, the findings indicate the need for systematic study of boundary-setting procedures, the robustness of grade mapping under small rank shifts, and the construct relevance of comparative judgments under realistic time constraints. Future research should therefore focus on identifying the conditions under which ACJ might be used more defensibly and practicably - for example, by testing ACJ as a calibration tool in authentic high-stakes settings, examining how different standard-setting procedures affect grade outcomes, and evaluating whether structured rationales or carefully governed AI-based supports can improve transparency without undermining validity and explainability requirements.

Data availability statement

The datasets presented in this article are not readily available because student examination scripts and full survey-response datasets cannot be shared due to confidentiality, institutional policy, and data protection requirements; only fully anonymized,

non-identifiable extracts/aggregates from the survey can be provided. De-identified RM Compare (ACJ) outputs, analysis code, and the boundary-setting protocol can be shared upon reasonable request. Requests to access the datasets should be directed to kjetil.egelandstal@uib.no.

Author contributions

KE: Writing – original draft, Writing – review & editing. J-OF: Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The study was funded by Centre for the Science of Learning and Technology (SLATE), University of Bergen.

Acknowledgments

The authors thank Siv Elén Årskog Vedvik, course coordinator for JUS2311 International Climate Law and Guri Lindblad, course coordinator for JUS2399 Tax Law for conducting the *post-hoc* grade calibration of the ACJ rank orders and for contributing to the structured discrepant-case review in their respective courses. They also facilitated access to the study context and provided valuable insight during the development of the project.

References

- Bartholomew, S. R., and Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *Int. J. Tech. Design Educ.* 32, 1159–1190. doi: 10.1007/s10798-020-09642-6
- Bloxham, S., and Boyd, P. (2007). *Developing effective assessment in higher education: A practical guide*. London: Open University Press.
- Boud, D. (2007). “Reframing assessment as if learning were important,” in *Rethinking assessment in higher education: Learning for the longer term*, eds. D. Boud and N. Falchikov (New York: Routledge) 14–25. doi: 10.4324/9780203964309
- Bramley, T., and Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assess. Educ. Principl. Policy Pract.* 25, 548–566.
- Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa
- Buckley, J., Canty, D., and Seery, N. (2020). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies*, 41. doi: 10.31235/osf.io/ys9eg
- Chambers, L., and Cunningham, E. (2022). Exploring the validity of comparative judgement: do judges attend to construct-irrelevant features? *Front. Educ.* 7:802392. doi: 10.3389/feduc.2022.802392
- Egelandsdal, K., Hartell, E., and Færstad, J. O. (2025). Exploring the practical feasibility of adaptive comparative judgment as a summative assessment method. *Assess. Eval. High. Educ.* 50, 1277–1292. doi: 10.1080/02602938.2025.2511787
- Hartell, E., and Buckley, J. (2021). “Comparative Judgment: An Overview,” in *Handbook for Online Learning Contexts: Digital, Mobile and Open: Policy and Practice*. eds. A. Marcus-Quinn and T. Hourigan. (Cham: Springer International Publishing) 289–307. doi: 10.1007/978-3-030-67349-9_20
- Jones, I., and Davies, B. (2024). Comparative judgement in education research. *International J. Res. Met. Educ.* 47, 170–181. doi: 10.1080/1743727X.2023.2242273
- Jones, I., and Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Stud. Educ. Eval.* 47, 93–101. doi: 10.1016/j.stueduc.2015.09.004
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Measure.* 50, 1–73. doi: 10.1111/jedm.12000
- Kelly, K. T., Richardson, M., and Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: a call for clarity. *Assess. Educ. Principl. Policy Pract.* 29, 674–688. doi: 10.1080/0969594X.2022.2147901
- Klenowski, V., and Wyatt-Smith, C. (2014). *Assessment for Education: Standards, Judgement and Moderation*. Thousand Oaks: SAGE. doi: 10.4135/9781526401878
- Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., and De Maeyer, S. (2022). Validity of comparative judgment scores: how assessors evaluate aspects of text quality when comparing argumentative texts. *Front. Educ.* 7:823895. doi: 10.3389/feduc.2022.823895
- Marshall, L., Shaw, S., Hunter, D., and Jones, A. (2020). Assessment by Comparative Judgement: An application to secondary statistics and English in New Zealand. *New Zealand J. Educ. Stud.* 55, 49–71. doi: 10.1007/s40841-020-00163-3
- McMahon, S., and Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assess. Educ. Principl. Policy Pract.* 22, 368–389. doi: 10.1080/0969594X.2014.978839

The authors further acknowledge the Faculty of Law, University of Bergen for enabling data collection, and the examiners who participated in the study.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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.

- Murphy, K. R., Jako, R. A., and Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *J. Appl. Psychol.* 78, 218–225. doi: 10.1037/0021-9010.78.2.218
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Princ. Policy Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354
- Pollitt, A., and Murray, N. L. (1993). *What Raters Really Pay Attention to*. Research Centre for English and Applied Linguistics. Cambridge: Cambridge University.
- RM (2024). RM Compare [Web-based adaptive comparative judgment platform]. Available online at: <https://compare.rm.com/> (Accessed March 4, 2026).
- Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assess. Eval. High. Educ.* 34, 159–179. doi: 10.1080/02602930801956059
- Steedle, J. T., and Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Appl. Measure. Educ.* 29, 211–223. doi: 10.1080/08957347.2016.1171769
- Steiner, D. D., and Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *J. Appl. Psychol.* 74, 136–142. doi: 10.1037/0021-9010.74.1.136
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288
- 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 den Berg, S., and Papadopoulos, P. M. (2024). Summative assessment with Artificial Intelligence: Qualitative analysis and comparison of technology acceptance in student and teacher populations. *Innov. Educ. Teach. Int.* 62, 1529–1544. doi: 10.1080/14703297.2024.2436613
- 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
- Xact by Ramboll. (2024). SurveyXact [Web-based survey software]. Available online at: <https://rambollxact.com/surveyxact> (Accessed March 4, 2026).