Article:
Chamberlain, J., Francis, S. and Herrick, T. (2026) Assessor experience, not rubric type, determines grading reliability in biosciences coursework. Frontiers in Education, 11. 1729644. ISSN: 2504-284X

https://doi.org/10.3389/feduc.2026.1729644

frontiers in Education

TYPE Original Research
PUBLISHED 22 April 2026
DOI 10.3389/feduc.2026.1729644

OPEN ACCESS

EDITED BY
Peter Ralph Grainger,
University of the Sunshine Coast, Australia

REVIEWED BY
Carolina Lopera-Oquendo,
The City University of New York, United States
Mardy SEREY,
Svay Rieng University, Cambodia

*CORRESPONDENCE
Janet Chamberlain
j.chamberlain@sheffield.ac.uk

RECEIVED 21 October 2025
REVISED 23 March 2026
ACCEPTED 30 March 2026
PUBLISHED 22 April 2026

CITATION
Chamberlain J, Francis S and Herrick T (2026) Assessor experience, not rubric type, determines grading reliability in biosciences coursework.
Front. Educ. 11:1729644.
doi: 10.3389/feduc.2026.1729644

Assessor experience, not rubric type, determines grading reliability in biosciences coursework

Janet Chamberlain[1]*, Sheila Francis[1] and Tim Herrick[2]

[1]Division of Clinical Medicine, School of Medicine and Population Health, University of Sheffield, Sheffield, United Kingdom, [2]School of Education, University of Sheffield, Sheffield, United Kingdom

Introduction: Previous research on whether holistic or analytical scoring rubrics yield higher reliability and validity has been mixed and inconclusive, with little of it based on empirical data. This study addressed this gap by comparing scores from assessors with varying experience levels who used both a holistic and an analytical scale descriptor to grade undergraduate bioscience essays, complemented by a qualitative survey of assessor perceptions. The goal was to determine which method resulted in more consistent inter-rater agreement.
Methods: For the study, independent assessors (two per essay) scored essays using either holistic ($n = 212$) or analytical ($n = 62$) scale descriptors, over four consecutive years. Each assessor provided both a holistic and an analytical score for every essay. The agreement between the scores was then calculated.
Results: The results showed no significant difference in scores awarded holistically versus analytically for the same essay, regardless of the scale descriptor used or the assessor's experience. However, a key finding was that experienced assessors had a higher agreement with the final awarded grade than their less-experienced counterparts.
Discussion: This suggests that the experience of the assessor, rather than the specific scoring rubric or guidance provided, is the primary factor in determining the reliability of the grade awarded.

KEYWORDS
analytic, assessment guidance, holistic, rubric, scoring

1 Introduction

Despite decades of research into grading (Brookhart et al., 2016), and the centrality of its position within academic practices, how it is understood in the assessment of higher education assignments, and how it is pursued in an equitable and transparent manner, remain questions that have only partially been resolved. One area where debate is ongoing is around scoring rubrics, defined as "a scoring tool for qualitative rating of authentic or complex student work" including "criteria for rating important dimensions of performance, as well as standards of attainment for those criteria" (Jönsson and Svingby, 2007). Rubrics hold the promise of greater consistency in the marking process, and improved transparency for students; at the same time, they risk reducing the performance of a complex set of academic tasks to a set number of observable traits, which sits poorly in relation to grander aims of higher education as a process of potential transformation. Scoring rubrics can take a holistic form, where the work is judged globally and a single scoring decision is made, or analytic, where it is judged over

several separate considered dimensions, each given an individual score that makes up a final grade. The validity and reliability of each scoring method has long been debated (Davis, 2018; Hamp-Lyons, 1995; Lee et al., 2009; Tomas et al., 2019; Weigle, 2002).

Holistic scoring is suited to situations where a quick decision is needed, and allow for sophisticated, higher-level skills, such as critical thinking, to be recognised (Tomas et al., 2019). Because this scoring is less granular in its approach, it appeals to prioritising the expert judgement of scorers, which is held by some to be a distinguishing feature of assessment in higher education given its intellectual complexity and the weight placed on the specialist knowledge of the scorers (Bloxham et al., 2016). Such scoring also appears to encourage scorers to make more explicit reference to the criteria in feedback, which is of benefit to learners (Jönsson et al., 2021).

Equally, holistic scoring can add challenges to interpreting the reasoning behind the score where two markers disagree and grade moderation is required (Brown, 1995; Xi, 2007). It can also be subject to marker bias, where a marker focuses on what a candidate does well to the detriment of areas of weaknesses (Bacha, 2001).

Analytic scoring, in comparison, makes it easier to interpret where scoring disagreements arise (Barkaoui, 2010) and can be beneficial to inexperienced scorers that may need detailed guidance to consider the wide range of scoring criteria, leading to greater reliability (Goulden, 1994). In addition, the determination of an overall grade from a collection of individual scores can lead to more consistent marking (Barkaoui, 2011), more regular reference to grades (Jönsson et al., 2021), and reveal strengths and weaknesses in an assignment to better aid feedback to a learner (Bacha, 2001). However, this approach also has disadvantages. Several (nominally independent) scoring decisions are required per assignment and this can lead to a "halo effect" with scorers finding it difficult to delineate between different aspects of an assignment and awarding similar marks in each category (Lai et al., 2015; Sheppard, 2019). This cognitive bias can also lead to a scorer being influenced by their positive or negative impression of one of the specific criteria, causing inconsistent or inaccurate scoring across the other, unrelated, criteria (Lai et al., 2015; Sheppard, 2019). The halo effect can be especially prevalent in cases where marking is done in one sitting and essentially leads to the markers making a holistic judgement overall. Additionally, if the weighted importance of each individual criteria is misaligned, the final mark may not be an accurate representation of the quality of the work (Tomas et al., 2019) and inexperienced scorers can have difficulty reliably distinguishing between criteria (Xi, 2007).

Whichever scoring rubric is used, there is a need for a reliable and detailed scale descriptor to aid in the decision making. A scale descriptor is understood as a qualitative description of what is expected at each grade point; to stop here would mean a scale descriptor holistic in its approach, whereas an analytical scale extends the differentiation by grade point into each aspect of an assignment. High quality scale descriptors find a suitable balance between detail of description, so scorers and students can easily relate their understanding of a submission to a grade, and flexibility of interpretation, to recognise the diversity of forms of response any assessment task might receive. Yet at the same time, this flexibility means that scale descriptors are subject to interpretation and consistent application needs training or familiarity with the assignment subject and descriptor definitions (Brookhart, 2018; Davis, 2018; Fulcher, 2003).

However effective the rubric, it still needs to be utilised by a marker, and this necessarily brings in elements of variability. One potentially relevant dimension might be the level of experience of the marker: Lopera-Oquendo et al. (Lopera-Oquendo et al., 2024) analyse the responses of trainee teachers to holistic and analytic grading scales. They conclude that the level of experience of the marker was less significant than other factors, such as gender and personality traits. Another variable might be the subject background of the marker, with one study (Möller et al., 2022) suggesting a closer fit between marker's expertise and the subject matter of the assessment leads to more robust relative ranking of assignments but has little effect on the grades awarded.

Both of these previous studies focus on humanities subjects, in stages of education prior to university. In contrast, our study focuses on scientific writing within higher education. Scientific writing may be seen as a particular case because the assessment can examine a range of criteria from factual accuracy to particular requirements of the scientific writing style [e.g., (Peat et al., 2013)]. Factual responses that are right or wrong form only one part of scientific writing, especially in longer-form writing where clear articulation of the ideas, processes, and assumptions is paramount. There is also little consideration given to the role that scorer preference has on the use of different rubrics (Chan and Luk, 2022): are some markers more attracted to one style than another, and if so, how does this affect their rubric use?

As a result, this study addressed several linked questions, centred around determining whether an analytic or holistic scoring rubric was more reliable at producing inter-scorer agreement when marking scientific coursework essays, where the scorers had a varied range in experience and knowledge of the subject assessed, and different preferences towards style of rubric utilised. It addressed three key questions within the specific context of scientific coursework in higher education: 1) do analytic or holistic scoring rubrics aid in reliable inter-scorer agreement? 2) do analytic or holistic scale descriptors influence reliability? 3) do scorer preferences towards scales affect how they score? The contribution it makes is towards understanding the nuances of this complex debate, and developing robust practices that are sustainable at scale and over time, in a climate of higher education where resources are increasingly constrained and transparency of decisions is of growing importance.

2 Methods

This study was granted ethical approval by the University of Sheffield Ethical Review Panel. Reference numbers 048414 and 065176.

2.1 Empirical study: determination of scoring variation using analytic and holistic scoring rubrics

To investigate the impact of marking experience, scoring method (rubric type), and marking guidance (scale descriptor)

on assessment quality, this empirical arm of the study employed a two-stage, within-subjects (scorer) experimental design. The design used parallel essay cohorts from different assessment years to maintain consistency across the longitudinal comparison. The first stage of the study focused on marks awarded using a holistic scale descriptor, where scorers were provided with general assessment criteria for each grade range, with the second stage using an analytical scale descriptor, providing scorers with a structured framework for evaluating specific criteria. To control for scorer-specific variance, the same pool of markers participated in both stages. This pool was divided into two distinct cohorts: experienced scorers and inexperienced scorers to enable the evaluation of the impact of professional experience on marking consistency and accuracy. To prevent confounding due to scorer-essay assignment, essays were allocated based on Subject Matter Expertise (SME), defined as holding a postgraduate qualification (PhD) in the specific sub-discipline. Markers were matched to specific essay titles to ensure they only evaluated content within their established field of expertise. This ensured that scorer judgements were grounded in content-specific knowledge rather than generalist impressions. Simultaneously, memory-based carryover effects were prevented by using different non-overlapping, essay cohorts for each scale descriptor.

Fairness was maintained through a Reference Rater protocol, where a single experienced scorer (the 'Anchor') assessed the entire cohort of essays across both cycles. Furthermore, all accuracy metrics (Weighted Kappa) were calculated against an independently-determined 'Real Grade' established by a separate moderation process for each year, thereby normalising for any inherent differences in essay difficulty between the assessment cycles.

The 'Real Grade' benchmark scores were established using the University's standard double-marking protocol. In accordance with institutional policy, at least 10% of the total cohort underwent formal moderation. This moderated sub-sample specifically targeted scripts where the initial marker discrepancy exceeded 5% or where marks fell near classification boundaries. For the remaining scripts where markers differed by fewer than 5 marks, the two scores were averaged to produce the final benchmark. Statistical comparisons in this study used the participants' initial raw scores against these finalised institutional benchmarks.

Individual essays ($n = 274$) from a final year undergraduate biosciences module, 2,000 words in length, were used in this study. These essays were chosen as they were easily accessible in a sufficient number for statistical analysis and required grading on both scientific content and accuracy as well as more general scholarly writing skills. The essays were divided into subsets, with each subset addressing a different, but related, title related to the teaching delivered within the biosciences module. As these essays were generated for an actual student assessment, the real grade benchmark scores were used as a standard for validity testing within this study. Scores were collated over a period of 4 years, generating scores for 212 individual essays using a holistic scale descriptor and 62 individual essays using an analytical scale descriptor. For this study, each individual essay was independently marked by two scorers, and their marks compared.

Seven different scorers graded the essays, of which three were considered "inexperienced" (had subject knowledge but no prior formal coursework essay marking experience), and four were considered "experienced" (subject knowledge and over 5 years of experience in marking coursework essays). Each essay was marked by a combination of either two experienced ($n = 150$, holistic scale descriptor, $n = 27$ analytical scale descriptor), or one experienced and one inexperienced ($n = 62$, holistic scale descriptor, $n = 35$ analytical scale descriptor), scorer. A combination of two inexperienced scorers was not included as all essays would be scored by at least one experienced scorer in a real situation.

To ensure the validity of the scores and minimise rater fatigue, an asynchronous marking protocol was employed. Scorers were provided with their assigned essay subsets and a two-week window to complete the assessment. Within this timeframe, scorers were permitted to schedule their own scoring sessions and determine the order of evaluation. This naturalistic approach was intended to reflect authentic institutional marking practices and allow scorers to maintain optimal cognitive focus, thereby enhancing the reliability of the resulting data.

Scorers were not given formal training prior to scoring, to ensure that the study could isolate the specific impact of the holistic and analytical frameworks. This approach prevented training from becoming a confounding variable, allowing for a direct comparison of the two methods. It specifically tested if the scale descriptors provided enough clarity to improve marking consistency without the need for additional instruction or group calibration. To establish a reliable benchmark for the untrained participants, an experienced scorer (the module lead) acted as an "anchor marker" and evaluated all 274 essays. These scores served as the "gold standard" against which the marks of the other participants were measured. The remainder of the scorers assessed a specific subset of the essays based on their subject knowledge.

This design allowed the study to precisely track how each scale descriptor or scoring rubric influenced the markers' accuracy relative to a professional standard. By using this expert baseline, the research could determine if the scale descriptor or rubric helped untrained markers achieve a level of consistency similar to that of an experienced examiner. This ensured that even without formal training, the performance of the markers could be evaluated against a stable and valid academic grade.

The scorers were provided with either a holistic ($n = 212$ essays) or an analytical ($n = 62$ essays) scale descriptor. The holistic scale descriptor was designed by the subject matter expert leading the module. This scale descriptor follows a standardised institutional format and has been refined over several years of active use in undergraduate assessment. The analytic scale descriptor was devised from the holistic descriptor and gave the same information as the holistic version, but the expectations for each section of word count, grammar, use of citations, synthesis, subject focus, content coverage and scientific accuracy was provided separately for each grade, rather than provided as holistic information separated by grade alone (see Supplementary Information). The weighting of these sub-criteria followed the University's standardised marking framework, which assigns 10% for presentation and style and 90% for understanding and content. This approach ensured that the scale descriptors were contextually authentic and aligned with

the actual grading environment of the study participants. The weighting for each section was not given to the scorers.

The scorers were asked to assign a holistic score to each essay. In addition, the scorers were also asked to assign breakdown scores for the essay using an analytic rubric. To minimise the risk of scorers artificially aligning their scores, the specific weightings derived from the University's standardised marking framework were withheld from the participants. This blinding was intended to isolate the implicit values of the scorers. By preventing explicit knowledge of the rubric's mathematical structure, the study could more accurately measure the 'expertise gap' between the scorers' global impression and the institutionally mandated weights.

Essays were scored for their word count, grammar, use of citations, synthesis, subject focus, content coverage and scientific accuracy using a 100 point scale. To further ensure the independence of the two marking methods, markers were asked to assign a holistic score based on their global impression of the essay before proceeding to the analytical breakdown. This procedural control was to prevent 'mathematical priming' and to allow for a statistically valid comparison between intuitive professional judgment and rubric-mediated assessment.

To maintain the independence of individual sub-criteria judgments used in analytical scoring, scorers provided raw marks for each analytical component. The composite analytical score for each essay was then calculated by the study lead using the pre-determined weighting for each separate category. This procedure ensured that the final analytical grade was a precise mathematical reflection of the sub-scores, free from any manual rounding or 'grade-adjusting' by the scorers themselves. The essay grade was determined by classifying the scores into 1st (above 70), 2i (60–69), 2ii (50–59), 3rd (40–49) or pass (35–40), depending on the score awarded.

2.2 Statistical analysis of empirical data

All analyses were conducted using SPSS (v29) software. Statistical interpretation criteria and thresholds used for the statistical analyses in this study are defined in the (Supplementary Section 1.4) (Cohen, 1988; Koo and Li, 2016; Landis and Koch, 1977; Meteyard and Davies, 2020).

To determine whether the data sets from the three years of holistic scale descriptor scoring could be combined for aggregate analysis, homogeneity of variance was assessed using Levene's test and equality of mean scores was evaluated via a one-way ANOVA. These tests were conducted to ensure that neither the score distribution nor the average scores differed significantly by year, thereby justifying the consolidation of the data into a single dataset.

To determine whether inexperienced scorers gained experience over the duration of the study, thereby introducing a potential confounder, absolute error (the absolute deviation between the rater score and the moderated 'Real Grade') was calculated. To investigate scorer maturation (defined as a systematic change in a scorer's stringency or consistency over time) (Wolfe et al., 2001) a Linear Mixed Model (LMM) was conducted with scorer ID as a random effect and assessment year as a fixed factor. Supplementary Table S1 shows the statistical specifications of the LMM.

To determine whether the analytical breakdown scoring matched the holistic scoring of each individual scorer (intra-rater agreement), for each essay marked, an intraclass correlation coefficient (ICC) (two-way mixed, absolute agreement), was performed.

To measure the degree of consensus (reliability) between markers (inter-rater reliability), ICC were calculated using a two-way random effects model for absolute agreement.

The accuracy (validity) of awarded marks was evaluated using Pearson Correlation Coefficients (r) comparing the awarded mark with the essay "real mark". A weighted Cohen's Kappa measure (assuming ordered categories) was used to determine the level of agreement between the scorers' essay grade classification (1st, 2i etc.) with the real essay grade awarded.

To identify systematic bias and interaction effects, Linear Mixed Models (LMM) were used. This model was used because it effectively handles the varied number of essays marked by different scorers and allows separation of the influence of individual scorer opinion from the actual scoring rubrics and scale descriptors used.

For the initial analysis, separate models were built for each type of scale descriptor. In these models, scorer experience and scoring rubric were treated as the primary factors to see if the type of rubric used (analytical vs. holistic) affected experienced and inexperienced scorers differently. To ensure the results were not skewed by the quality of the student work itself, the 'real mark' was included as a 'control' variable (covariate). Finally, we included a random effect for Scorer ID; this essentially mitigates the fact that some scorers are naturally harsher or more lenient than others, allowing for a clearer view of how the rubrics themselves actually perform.

After investigating each scale descriptor individually, we used a more comprehensive three-way model to determine how scorer experience, scale descriptor type and scoring rubric type interacted together. This tested whether changing from holistic to analytical scale descriptors changed how experienced and inexperienced scorers used the scoring rubrics.

For ease of comparison, Estimated Marginal Means (EM means) were calculated, providing 'adjusted' scores that assumes every scorer graded essays of the same quality. As multiple comparisons were performed, a Bonferroni adjustment was applied to all *post-hoc* pairwise comparisons as a statistical safeguard against random chance. This overall strategy ensured the data were robust and not skewed by individual scorer bias or a particularly difficult set of essays.

2.3 Determination of staff preference for scoring and scale descriptors

To capture staff perceptions of the two scoring methods, an online qualitative survey was distributed to the departmental teaching team via a central institutional email alias. This resulted in 30 completed responses (a 32% response rate).

The survey focused on staff preferences regarding the two scale descriptor types and the perceived impact of each marking method on marking efficiency and justification. The resulting

qualitative data were analysed by thematic analysis conducted by a single coder using NVivo software following the six-phase framework established by Braun and Clarke (Braun and Clarke, 2006; Braun and Clarke, 2019): data familiarisation, the generation of initial open codes, the searching for and reviewing of broader themes, the definition and naming of final themes, and the final reporting. To ensure the credibility of the analysis, a systematic coding audit was maintained, and emerging themes were cross-referenced against the raw survey responses to ensure a faithful representation of staff preferences.

Although the analysis was conducted by a single coder, rigorous measures were employed to ensure the "trustworthiness" and "dependability" of the findings. This included the maintenance of a digital audit trail within NVivo and a process of peer debriefing, wherein the emerging thematic structure was reviewed to ensure that all interpretations were firmly grounded in the raw respondent data. The full survey used is provided in the Supplementary Material.

It is noted that while the participant groups for the quantitative and qualitative protocols overlapped, they were not identical, and individual preferences were not tracked to maintain participant anonymity. However, the qualitative study was specifically designed to capture the professional rationale behind the marking patterns observed in the empirical study. This integrated focus is a novel element of the study, allowing for the triangulation of actual scoring behaviour against perceived preference, for example, by revealing whether specific rubric formats function as genuine drivers of reliability or primarily as a psychological scaffold for scorers of varying experience.

3 Results

3.1 Descriptive statistics and data validation

Supplementary Table S2 presents the descriptive statistics (Mean, SD, and N) for the experienced and inexperienced scorers across the four-year study period.

Before analysing the experimental variables, we assessed the stability of the marking groups over the study period to ensure pooling the stage 1 data into single datasets for each group (experienced holistic scoring, experienced analytical scoring etc) was valid, and to determine whether inexperienced scorers gaining experience over the study duration was a confounding factor.

To ensure the datasets from the three-year data collection period using a holistic scale descriptor could be pooled for analysis, the assumption of homogeneity of variance was tested. Levene's test was non-significant [$F(2,134) = 1.24$, $p = 0.292$], indicating that marking score variance was consistent across the three years. A one-way ANOVA was then conducted to compare the effect of collection year on holistic scores. Results indicated no significant difference in scores between the three years [$F(2,134) = 0.704$, $p = 0.496$]. These findings suggest that the timing of data collection did not systematically bias scoring outcomes, justifying the pooling of the data for further analysis.

To assess potential scorer maturation, a Linear Mixed Model (LMM) was conducted on the absolute error scores for the inexperienced cohort across years 1–3. Analysis was restricted to this period to control for the change of scale descriptor introduced in year 4. A significant main effect of year was found for both holistic [$F(2, 17.17) = 11.79$, $p < 0{,}001$] and analytical [$F(2, 59) = 12.73$, $p < 0.001$] scoring. However, *post-hoc* analysis (Supplementary Table S3) revealed that this significance was attributable to inter-annual fluctuation rather than a linear maturation trend. Specifically, the Mean Absolute Error (MAE) rose significantly in year 2 (MAE = 9.13) compared to year 1 (MAE = 2.47–2.78), before decreasing again in year 3 (MAE = 3.73–4.44). This non-linear variation indicates that the significance was driven by annual cohort variance rather than a systematic improvement in scorer proficiency.

3.2 Scorer and reliability and systematic bias

To evaluate the internal consistency and potential bias within the marking process, initial analysis of intra-scorer agreement (comparing individual scorers' holistic vs. analytical marks) was undertaken, followed by an assessment of inter-rater reliability.

A comparison of the score awarded using the analytic scoring rubric with that of the holistic rubric for the same essay, by each individual scorer, by intra-class correlation analysis showed a very strong agreement between marks for both experienced and inexperienced scorers when using either a holistic scale descriptor or an analytical one (Table 1, visual representation shown in Figure 1). This suggests that neither the type of scoring rubric nor the type of scale descriptor used affects the score awarded by individual scorers.

When comparing the scores awarded by experienced and inexperienced scorers for individual essays, agreement was poor when using a holistic scale descriptor, regardless of whether a holistic or analytical scoring rubric was applied (Table 2; Figure 2). However, when the analytical scale descriptor was used, the level of agreement remained poor under the holistic scoring condition, but increased to a moderate level if analytical scoring was used.

When comparing scores awarded by two different experienced markers for individual essays, the agreement between scores was also poor for both types of scoring rubric when using a holistic scale descriptor. This agreement increased to a moderate level when using an analytical scale descriptor although there remained no difference between scoring method. (Table 2) (Supplementary Figure S1).

To determine if these reliability levels were influenced by systematic bias, we then carried out a series of Linear Mixed Models (LMMs).

Firstly, two-way LMMs were conducted for each scale descriptor type separately to isolate the interaction between scorer experience and scoring rubric. These subset models showed a high degree of systemic stability, with no significant interaction between scorer experience and scoring rubric when using a holistic scale descriptor (Supplementary Table S4). In contrast, when using an analytical scale descriptor, a significant effect was seen for the experience of the scorer ($t = 6.782$, $p < 0.001$). Pairwise comparison of mean marks awarded showed experienced scorers awarded significantly lower marks than

TABLE 1 Two Way mixed method, absolute agreement ICC: looking at intra-marker agreement.

Rubric comparison	Marking experience	ICC	95% Confidence interval	*F*	*df1*	*df2*	*p*	Correlation interpretation
Holistic guidance								
Analytic vs. holistic	experienced	0.926	[0.88, 0.95]	32.15	211	211	<0.001	Excellent
Analytic vs. holistic	inexperienced	0.978	[0.93, 0.99]	4.01	61	61	<0.001	Excellent
Analytical guidance								
Analytic vs. holistic	experienced	0.962	[0.92, 0.98]	20.1	82	82	<0.001	Excellent
Analytic vs. holistic	inexperienced	0.968	[0.91, 0.99]	10.8	32	32	<0.001	Excellent

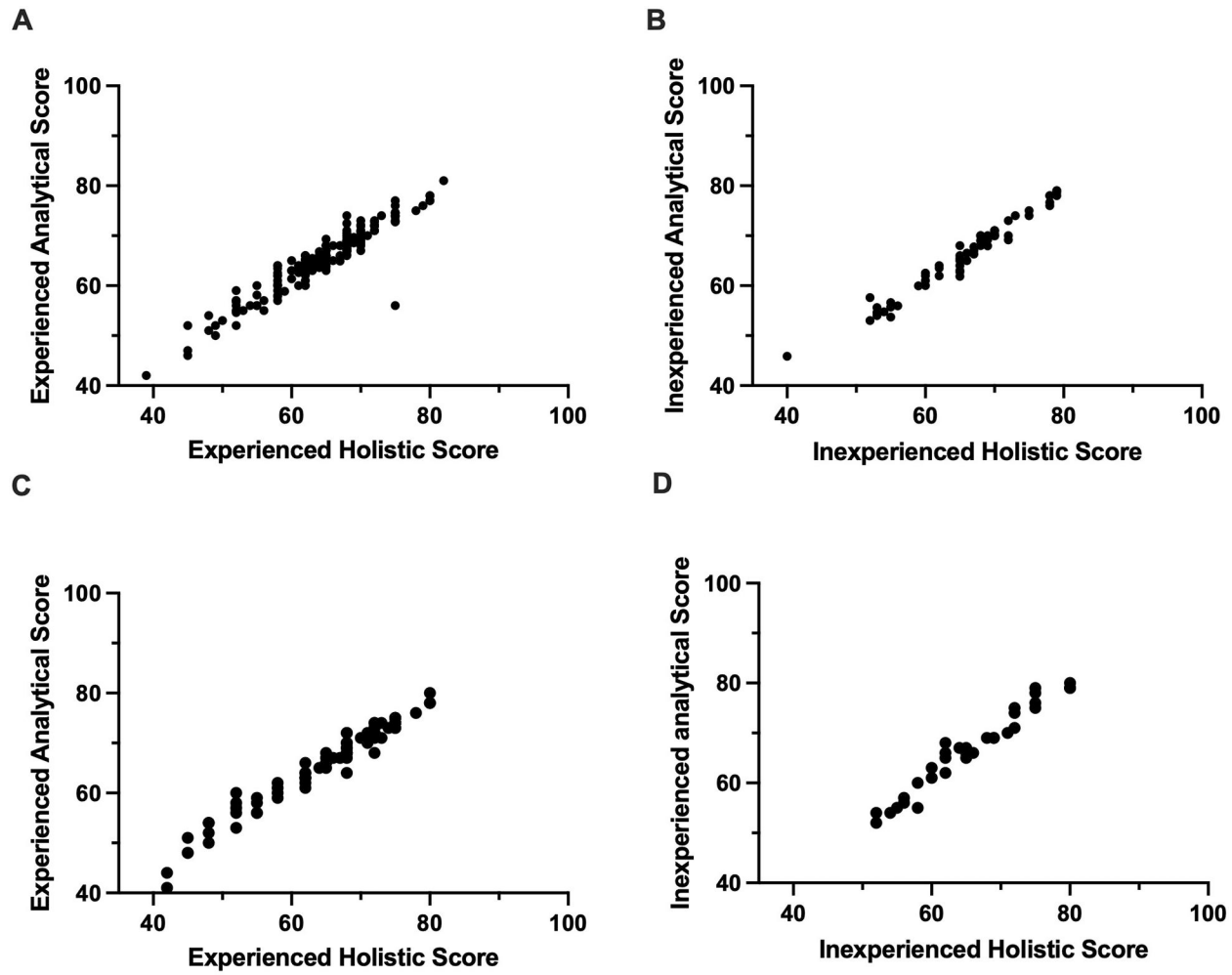

FIGURE 1
Intra-scorer correlation between holistic and analytical scoring methods. Graphs depict the relationship between scores awarded to the same essay by individual experienced **(A,C)** and inexperienced **(B,D)** scorers using either holistic **(A,B)** or analytical **(C,D)** scale descriptors. Each data point represents a single essay; the distribution along the axes reflects the range of essay quality within the sample. The strong positive correlations indicate high internal consistency for individual scorers regardless of the descriptor used.

inexperienced scorers across both scoring rubrics (Mean difference = 2.535, $p < 0.001$).

A comprehensive three-way factorial LMM was then used to evaluate the overarching relationship between scale descriptor type, experience, and scoring rubric. The results (Table 3), showed significant main effects for both scorer experience ($p = 0.006$) and scale descriptor ($p = 0.010$), while the scoring rubric (holistic vs. analytical) did not significantly influence awarded scores ($p = 0.541$). Most notably, a significant two-way interaction emerged between experience and scale descriptor ($p = 0.035$).

The significant interaction between experience and scale descriptor was supported by the estimated marginal means, (adjusted for the real mark covariate) (Table 4). Specifically, the

TABLE 2 Intraclass correlation coefficients (ICC) for marker pairs by scoring method and scale descriptor.

Scorer pair	Marking method	ICC	95% CI	*F*	*df1*	*df2*	*p*	Correlation interpretation
Holistic guidance								
Exp vs. exp	holistic	0.489	[0.297, 0.643]	3.02	74	74	<0.001	Poor
Exp vs. exp	analytical	0.490	[0.287, 0.651]	2.96	66	66	<0.001	Poor
Exp vs. inexp	holistic	0.432	[0.209,0.612]	2.55	61	61	<0.001	Poor
Exp vs. inexp	analytical	0.425	[0.200, 0.608]	2.49	61	61	<0.001	Poor
Analytical guidance								
Exp vs. exp	holistic	0.646	[0.339, 0.828]	4.50	24	24	<0.001	Moderate
Exp vs. exp	analytical	0.641	[0.332, 0.825]	4.44	24	24	<0.001	Moderate
Exp vs. inexp	holistic	0.439	[0.128, 0.674]	2.80	32	32	0.002	Poor
Exp vs. inexp	analytical	0.550	[0.254, 0.750]	3.82	32	32	<0.001	Moderate

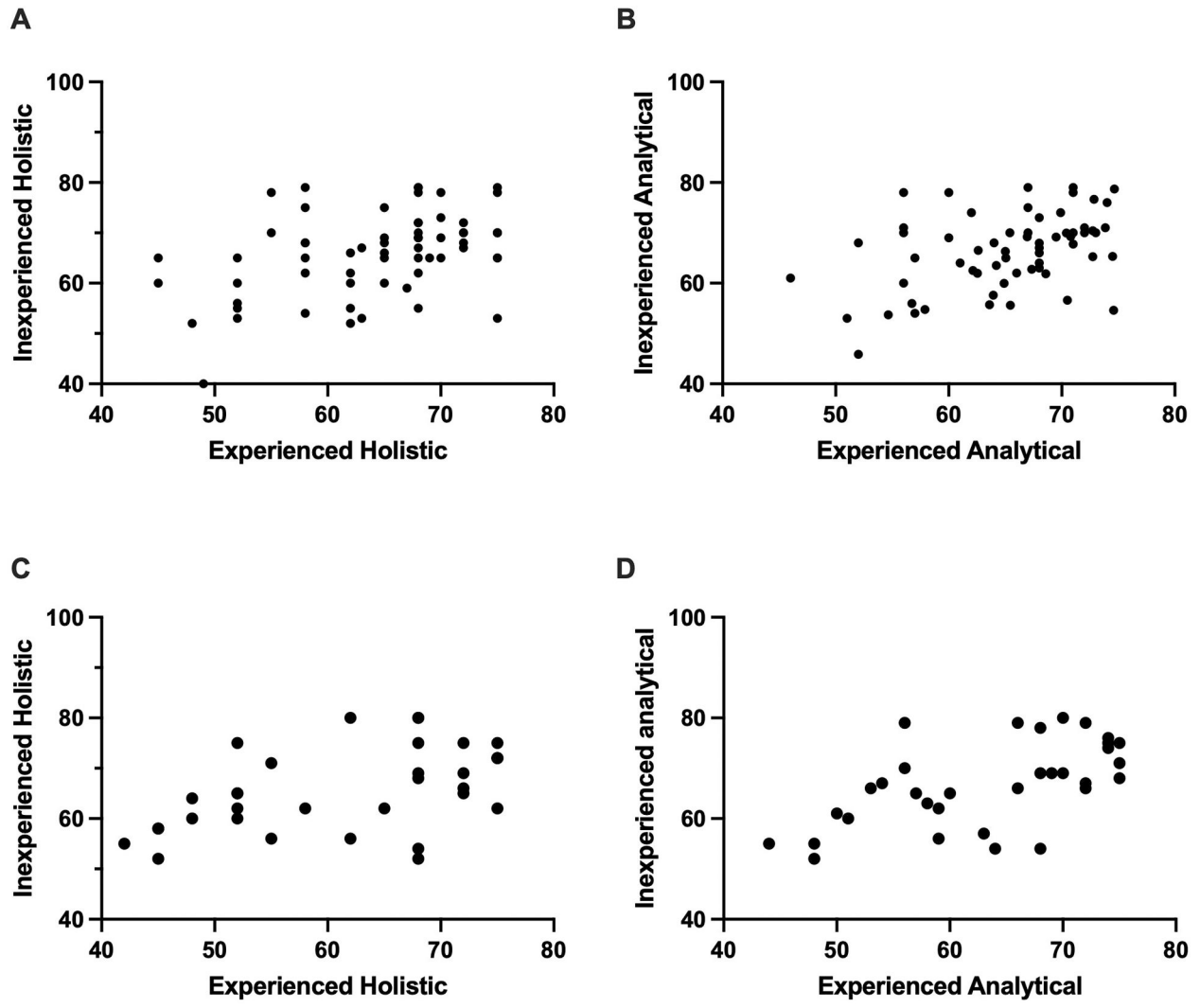

FIGURE 2
Correlation of scores awarded by inexperienced compared to experienced scorers using either holistic **(A,C)** or analytical **(B,D)** scoring rubrics and holistic **(A,B)** or analytical **(C,D)** scale descriptors. A poor agreement was seen using a holistic scale descriptor for both scoring rubrics. When using an analytical scale descriptor, poor agreement was seen using holistic scoring but this increased to give a moderate agreement if using analytical scoring.

TABLE 3 3-way linear mixed model results for marking accuracy across experience, method, and guidance.

Predictor (fixed effects)	Estimate (*b*)	*SE*	*df*	*t*	*p*
Intercept	12.471	1.382	853	9.026	<0.001
Real mark (covariate)	0.834	0.020	853	42.027	<0.001
Experience	−1.647	0.603	853	−2.732	0.006
Scoring rubric	0.468	0.765	853	0.611	0.541
Scale descriptor	2.334	0.920	853	2.537	0.010
Experience × rubric	0.639	0.854	853	0.749	0.454
Experience × descriptor	−2.248	1.065	853	−2.110	0.035
Rubric × descriptor	0.396	1.299	853	0.305	0.761
Experience × rubric × descriptor	−0.334	1.506	853	−0.222	0.825
Predictor (random effects)	**Variance**	*SD*	**Wald *Z***	*p*	
Marker (intercept)	23.091	4.805	20.761	<0.001	
Residual	52.158	7.222			

Model Fit Indices: AIC = [4,967.313], BIC = [5,019.665], $n = 870$ observations nested within $n = 8$ markers.

TABLE 4 Estimated marginal means for awarded marks by experience and marking condition. .

Experience	Scoring rubric	Scale descriptor	Mean (EMM)	Std error (SE)
Experienced	Holistic	Holistic	64.35	0.266
	Holistic	Analytical	64.44	0.468
	Analytical	Holistic	65.46	0.270
	Analytical	Analytical	65.61	0.468
Inexperienced	Holistic	Holistic	66.00	0.541
	Holistic	Analytical	68.33	0.744
	Analytical	Holistic	66.47	0.541
	Analytical	Analytical	69.20	0.743

Covariates appearing in the model are evaluated at the following values: Real essay score = 64.15.

transition from using a holistic rubric and holistic scale descriptor to a pure analytical framework resulted in an inflation of adjusted marks for the inexperienced cohort, rising from 65.99 to 69.19. The gap between experienced and inexperienced scorers was most narrow when using a holistic scale descriptor (1.64 points) and widest when using analytical rubrics (3.59 points).

In contrast, all interactions involving the scoring rubric were non-significant ($p > 0.05$) and the real mark remained a highly significant predictor of awarded scores ($t = 42.027$, $p < 0.001$).

The number of observations per experimental group used in the comprehensive three-way Linear Mixed Model are given in Supplementary Table S5.

To investigate why the global models showed such variance, ICC and MAE were calculated for the individual sub-scores used in the analytical rubric (Table 5). Using a holistic scale descriptor, ICCs were significant ($p < 0.05$) but demonstrated poor correlation (ICC < 0.50) across all except the synthesis category between two experienced scorers. Reliability failed to improve when using the analytical scale descriptor, where several ICCs actually became non-significant ($p > 0.05$). Furthermore, the Mean Absolute Error remained consistently high and did not decrease as scorers moved from holistic to analytical tools.

3.3 Inter-scorer accuracy and criterion validity

While the LMM showed that scoring rubrics had no significant effect on the average mark, it highlighted that scorers were influenced by the scale descriptors used. However, these findings do not indicate the actual accuracy of the scores. Therefore, the marks awarded were compared to the real marks of the essays to determine which conditions facilitated the most accurate scorer alignment.

Pearson correlation coefficients (r) showed that experienced scorers maintained a very strong correlation with the real essay scores ($r = 0.85$–0.93) regardless of the scoring rubric or scale descriptor employed (Table 6) (Visual representation in Figure 3). In contrast, inexperienced scorers demonstrated a

TABLE 5 Intraclass correlation and mean absolute error for individual Sub-scores used in the analytical scoring rubric.

Marker pair	Criteria	MAE	ICC	95% CI	*F*	*df1*	*df2*	*p*	Correlation
Holistic scale descriptor									
Exp vs. Exp	Word count	5.17	0.299	[0.081, 0.492]	1.950	70	70	0.003	Poor
	Grammar	5.99	0.160	[−0.071, 0.375]	1.386	70	70	0.087	Poor
	Citation use	5.28	0.454	[0.249, 0.621]	2.656	70	70	<0.001	Poor
	Synthesis	5.99	0.513	[0.320, 0.665]	3.115	70	70	<0.001	Moderate
	Focus	5.90	0.420	[0.212, 0.592]	2.536	70	70	<0.001	Poor
	Coverage	6.13	0.393	[0.182, 0.571]	2.335	70	70	<0.001	Poor
	Accuracy	6.07	0.429	[0.219, 0.600]	2.500	70	70	<0.001	Poor
Exp vs. Inexp	Word count	6.11	0.514	[0.305, 0.677]	3.288	61	61	<0.001	Moderate
	Grammar	4.82	0.361	[0.127, 0.588]	2.144	61	61	0.002	Poor
	Citation use	7.00	0.416	[0.186, 0.602]	2.411	61	61	<0.001	Poor
	Synthesis	6.69	0.337	[0.099, 0.539]	2.022	61	61	0.003	Poor
	Focus	6.50	0.375	[0.139, 0/571]	2.187	61	61	<0.001	Poor
	Coverage	6.56	0.404	[0.180, 0.595]	2.387	61	61	<0.001	Poor
	Accuracy	6.73	0.383	[0.147, 0.577]	2.222	61	61	<0.001	Poor
Analytic scale descriptor									
Exp vs. Exp	Word count	14.24	0.041	[−0.127, 0.283]	1.181	24	24	0.344	Poor
	Grammar	11.08	−0.014	[−0.281, 0.312]	0.964	24	24	0.535	Poor
	Citation use	5.68	0.721	[0.467,0.866]	6.169	24	24	<0.001	Moderate
	Synthesis	6.40	0.459	[0.097, 0.717]	2.734	24	24	0.008	Poor
	Focus	5.60	0.568	[0.238, 0.782]	3.660	24	24	<0.001	Moderate
	Coverage	7.04	0.399	[0.016, 0.681]	2.323	24	24	0.022	Poor
	Accuracy	6.28	0.466	[0.112, 0.720]	2.853	24	24	0.006	Poor
Exp vs. Inexp	Word count	12.88	0.246	[−0.068, 0.553]	2.282	32	32	0.011	Poor
	Grammar	6.67	0.329	[−0.016, 0.603]	1.958	32	32	0.031	Poor
	Citation use	6.94	0.705	[0.342, 0.864]	7.810	32	32	<0.001	Moderate
	Synthesis	8.00	0.527	[0.236, 0.733]	3.446	32	32	<0.001	Moderate
	Focus	7.27	0.395	[0.062, 0.648]	2.283	32	32	0.011	Poor
	Coverage	10.82	0.358	[0.043, 0.614]	2.253	32	32	0.012	Poor
	Accuracy	8.91	0.483	[0.181, 0.704]	2.930	32	32	0.002	Poor

TABLE 6 Pearson correlation for scorers compared to the real essay score, by rubric type and scale descriptor type.

Scorer pair	Scoring rubric	*r*	*n*	95% Confidence interval	*df*	*p*	Correlation interpretation
Holistic scale descriptor							
Experienced vs. real mark	Holistic	0.878	212	[0.843, 0.906]	210	<0.001	Very strong
Experienced vs. real mark	Analytical	0.854	204	[0.812, 0.887]	202	<0.001	Very strong
Inexperienced vs. real mark	Holistic	0.656	62	[0.486,0.778]	60	<0.001	Good
Inexperienced vs. real mark	Analytical	0.636	62	[0.459, 0.764]	60	<0.001	Good
Analytical Scale Descriptor							
Experienced vs. real mark	Holistic	0.937	83	[0.904, 0.959]	81	<0.001	Very strong
Experienced vs. real mark	Analytical	0.931	83	[0.895, 0.955]	81	<0.001	Very strong
Inexperienced vs. real mark	Holistic	0.632	33	[0.369, 0.801]	31	0.002	Good
Inexperienced vs. real mark	Analytical	0.644	33	[0.386, 0.809]	31	<0.001	Good

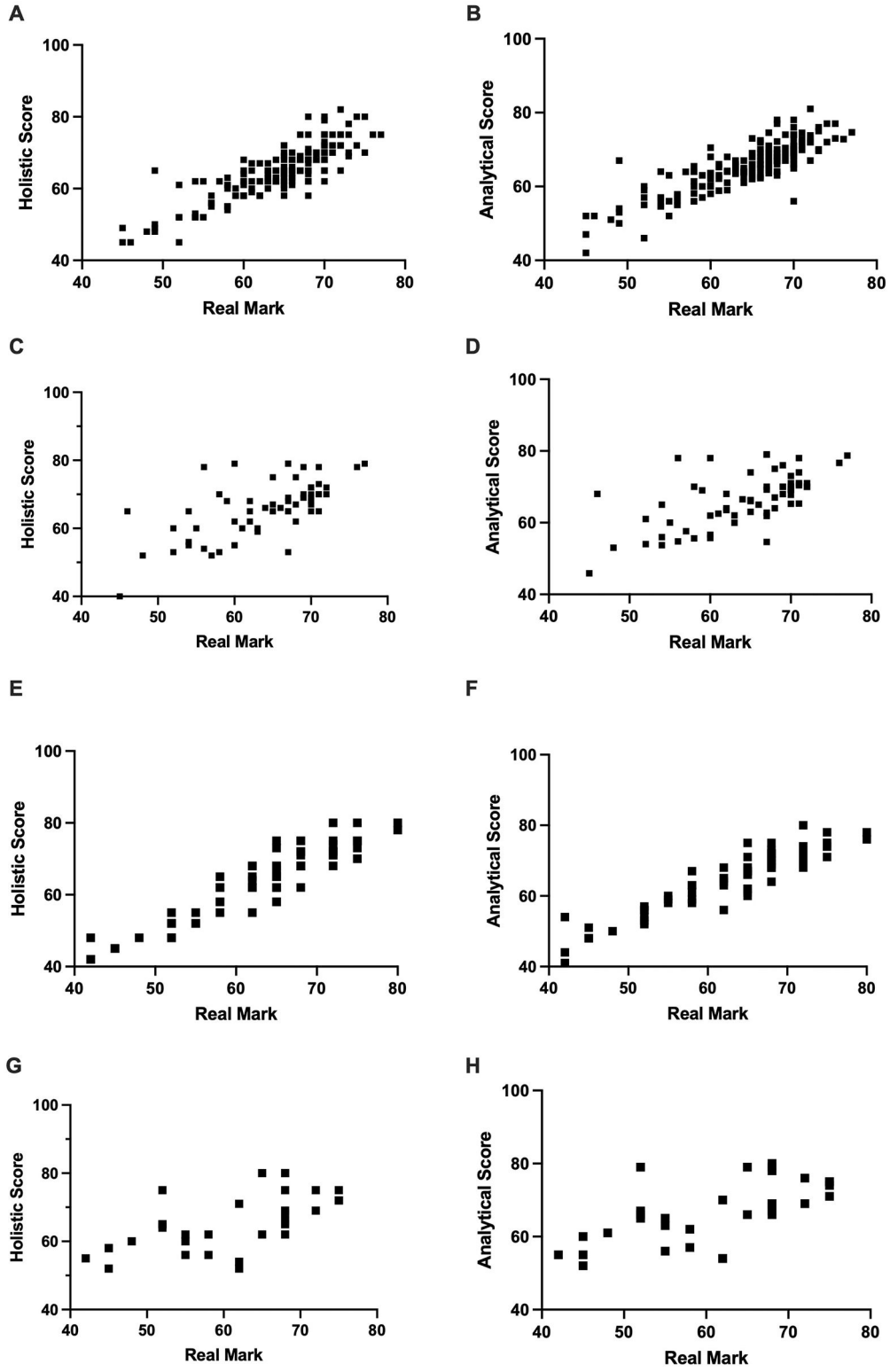

FIGURE 3
Correlation of real essay score and scores awarded using holistic **(A,C,E,G)** and analytical **(B,D,F,H)** scoring rubrics and holistic **(A–D)** or analytical **(E–H)** scale descriptors by experienced and inexperienced scorers. Experienced scorers **(A,B,E,F)** have a strong correlation with the real score, regardless of rubric used. Inexperienced scorers **(C,D,G,H)** have a positive, but less strong correlation than experienced scorers. The type of scale descriptor has no effect on these correlations.

good correlation ($r = 0.63–0.66$), although these levels were similarly consistent irrespective of rubric and scale descriptor.

Given that 100-point raw scores are ultimately translated into degree classifications for official record, a linearly weighted Cohen's Kappa (*Kw*) was calculated to assess the categorical agreement between the scorers assigned grade classifications (1st, 2i etc.) and the real grade classification.

TABLE 7 Weighted Cohen's kappa for scorers compared to the real essay grade, by rubric type and scale descriptor type.

scorer pairing	Scoring rubric	K_w	95% CI	p	K_w	95% CI	p
		Holistic guidance			Analytical guidance		
Exp vs. real	Holistic	0.793	[0.73, 0.86]	<0.001	0.890	[0.83, 0.95]	<0.001
Exp vs. real	Analytical	0.751	[0.68, 0.83]	<0.001	0.848	[0.78, 0.92]	<0.001
Inexp vs. real	Holistic	0.616	[0.44, 0.80]	<0.001	0.525	[90.34, 0.71]	<0.001
Inexp vs. real	Analytical	0.601	[0.42, 0.78]	<0.001	0.481	[0.30, 0.66]	<0.001
Exp vs. Exp	Holistic	0.453	[0.23, 0.68]	<0.001	0.667	[0.38, 0.95]	<0.001
Exp vs. Exp	Analytical	0.435	[0.21, 0.66]	<0.001	0.729	[0.53, 0.92]	<0.001
Exp vs. Inexp	Holistic	0.388	[0.17, 0.61]	<0.001	0.407	[0.20, 0.61]	<0.01
Exp vs. Inexp	Analytical	0.411	[0.19, 0.64]	<0.001	0.479	[0.28, 0.68]	<0.01

TABLE 8 Thematic summary of scorer perceptions regarding rubric utility and scale descriptor interpretation.

Overarching theme	Core sub-themes	Narrative summary
Structural support & transparency	Aids Expectations ($n = 21$); Aids Feedback ($n = 17$); Fairer Grading ($n = 16$); students see progress ($n = 6$)	Participants identified the analytical scale descriptor as a primary driver of transparency, noting it "anchored" expectations and provided a defensible framework for student feedback.
Reliability vs. Flexibility	Consistency ($n = 21$); Reliability ($n = 7$); Flexibility ($n = 12$); Hard to be Fair ($n = 11$).	A functional trade-off emerged between the perceived reliability of analytical scoring and the "flexibility" of holistic approaches. Holistic scoring was valued for its speed but criticised as "hard to be fair" without deep subject familiarity.
System constraints & "Rule-breaking"	Inflexible rubric ($n = 15$); Rule-breaking ($n = 10$); Bias ($n = 10$); Grade adjustment ($n = 42$).	When analytical rubrics were perceived as "inflexible" or "flawed", scorers admitted to "rule-breaking" and manual grade overrides. This suggests that expert intuition often supersedes the calculated score of a granular rubric.
Challenges to evaluative clarity	Unclear descriptors ($n = 10$); Section guides ($n = 5$); Assignment familiarity ($n = 3$).	Regardless of rubric type, the "unclear" nature of specific scale descriptors remained a barrier. Participants noted that "section guides" were helpful but could not replace the need for professional experience.

Experienced scorers showed a substantial agreement ($Kw = 0.75$–0.89, $p < 0.001$) with the real awarded grade, regardless of scale descriptor or scoring rubric used (Table 7). In contrast, inexperienced scorers showed only moderate agreement throughout ($Kw = 0.48$–0.61, $p < 0.001$). Notably, while agreement for this cohort remained moderate across all conditions, holistic scale descriptors yielded slightly higher correlation than analytical ones.

When examining inter-rater reliability between independent scorers, experienced scorers achieved substantial agreement only when using analytical scale descriptors, with the level of agreement between scorers falling to a moderate level using a holistic scale descriptor. Conversely, pairings consisting of one experienced and one inexperienced scorer yielded no higher than moderate agreement, though analytical descriptors facilitated a closer alignment between these cohorts.

3.4 Qualitative results: thematic analysis of scorer perceptions

To contextualise the quantitative findings, a thematic analysis was conducted on the free-text responses from 30 participants (19 experienced, 7 reasonably experienced, and 4 inexperienced) regarding rubric and scale descriptor preference. Following iterative coding in NVivo, an initial 32 sub-codes were synthesised into four overarching themes (see Table 8). The analysis reveals a clear tension: while participants valued the structural support of analytical tools for feedback purposes, they frequently encountered system constraints that led to "rule-breaking" and manual grade adjustments to maintain what they perceived as professional marking standards, with experienced scorers more likely to perceive constraints than inexperienced scorers.

3.4.1 Theme 1: structural support and feedback

The majority of scorers (87%) prefer analytical scale descriptors (Figure 4). Thematic analysis suggests this is driven by the clarity these descriptors provide for individual components of an assignment, which allows for a more objective grade. One participant noted that this structure "*forces me as a marker to look beyond fluency of expression, and to really focus on content,*" while an inexperienced scorer added, "*I find the table easier to be consistent with as I only need to consider one aspect of it at once.*"

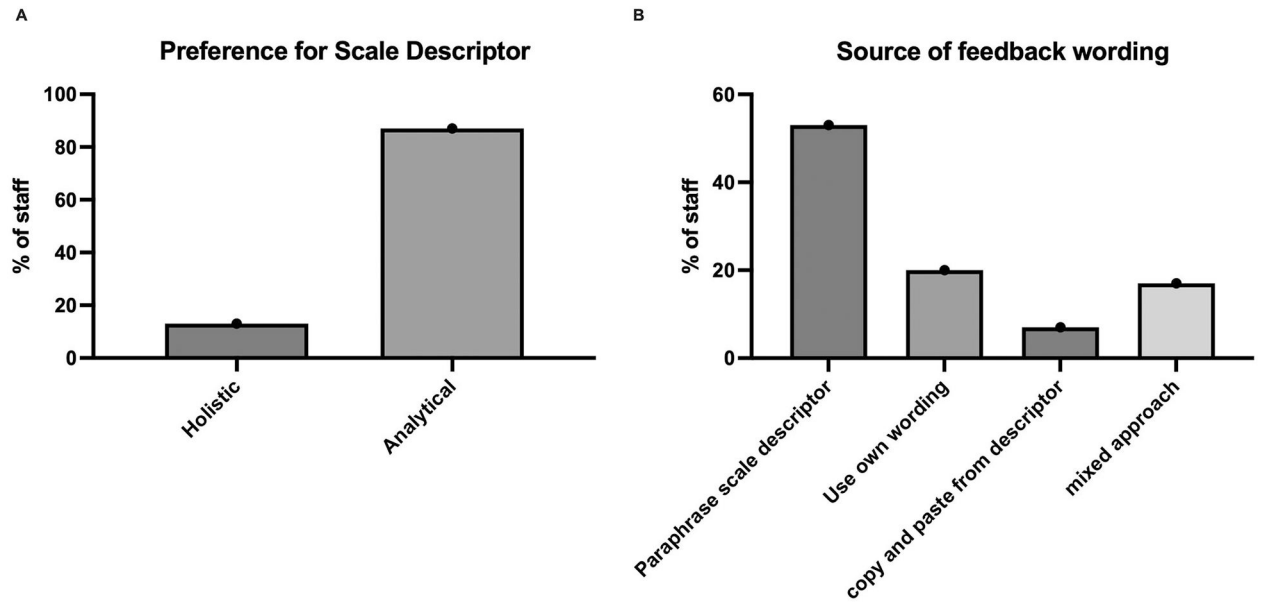

FIGURE 4
The majority of staff prefer to use an analytic scale descriptor **(A)**, and use this to inform the wording of their feedback **(B)**.

In contrast, only 13% of scorers expressed a preference for holistic scale descriptors (Figure 4). Opinions were mixed on the use of these, with some finding the descriptors easier to follow: "*I find the holistic method easier to understand and follow*" and others finding them difficult to ensure consistent marking: "*I think it is very difficult to standardise a holistic rubric as everybody is different in their marking style*".

Many scorers use an analytical scale descriptor to aid in writing feedback (Figure 4). Of these 53% paraphrase the rubric when writing feedback, 20% fully write feedback in their own words, and 27% use a mixed approach. Several scorers highlighted the positive aid that analytical scale descriptors have in giving feedback (17 mentions), allowing for structured feedback based on all aspects of the assignment, which gives clarity to the student on their strengths and weaknesses: "*The reason why I prefer an analytical scheme for writing feedback is that it can help flag where the assignment is overall to a student, even if it is a section that I don't have any major issues with.*" Only one scorer highlighted a negative experience with using an analytical scale descriptor with respect to writing feedback: "*I sometimes find the level descriptors of rubrics don't really fit with the kind of wording of feedback I would prefer to provide*"

When considering scoring rubrics, however, there was a belief that analytical scoring can have a negative effect on students with respect to feedback: "*students can become fixated on just the "bit that was bad" instead of considering the work as a whole and their work more globally.*"

3.4.2 Theme 2: reliability vs. flexibility of scoring

The survey showed staff prefer to mark analytically (57%) (Figure 5), although 33% of respondents have a different preference for scoring rubric, depending on the type of assignment being marked (Figure 5). Experienced scorers have a stronger preference for analytic scoring if they are unfamiliar with the assignment type or content, whilst inexperienced scorers tend to prefer analytical scoring rubrics overall.

Thematic analysis showed that scorers consider analytical scoring to lend itself to more intra-scorer consistency (21 mentions) but that this depends on correct weighting and explicit marking criteria. One participant stated this leads to more clarity: "*you arrive at a mark in a more transparent way*" with another highlighting the consistency: "*I prefer to mark analytically to make sure that my weighting of different components is comparable between students.*".

Several scorers also believed that analytical marking leads to less variance between scorers: "*Analytical marking helps to get better consistency in marking when there are multiple markers on a module*", and one highlighted that the transparency of marking breakdown allows for easier moderation where there is a disagreement between scorers: "*Analytical marking also helps when double marking or moderation—one examiner may have been a little too harsh or generous for a particular criterion so the markers can easily identify exactly where their marks may differ.*"

Although most respondents thought analytical scoring is more reliable, several expressed a preference for holistic scoring based on the flexibility to decide a grade (12 mentions): "*it isn't possible to design an analytical marking rubric that would adequately take account of everything a student might/should include*", believing this to be more "trustworthy": "*the overall mark when calculated from individual components, doesn't always reflect the piece of work as a whole and this is where holistic marking is sometimes better*".

Those that had concerns over holistic scoring (12 mentions) mostly highlighted fairness and bias as their reasons "*I do find that I over-estimate the value of fluent writing when marking holistically, and have to consciously force myself to be aware of this bias*", with one inexperienced scorer finding such scoring difficult to decide the grade: "*I find trying to use the whole grade structure might make it difficult to decide on individual marks.*"

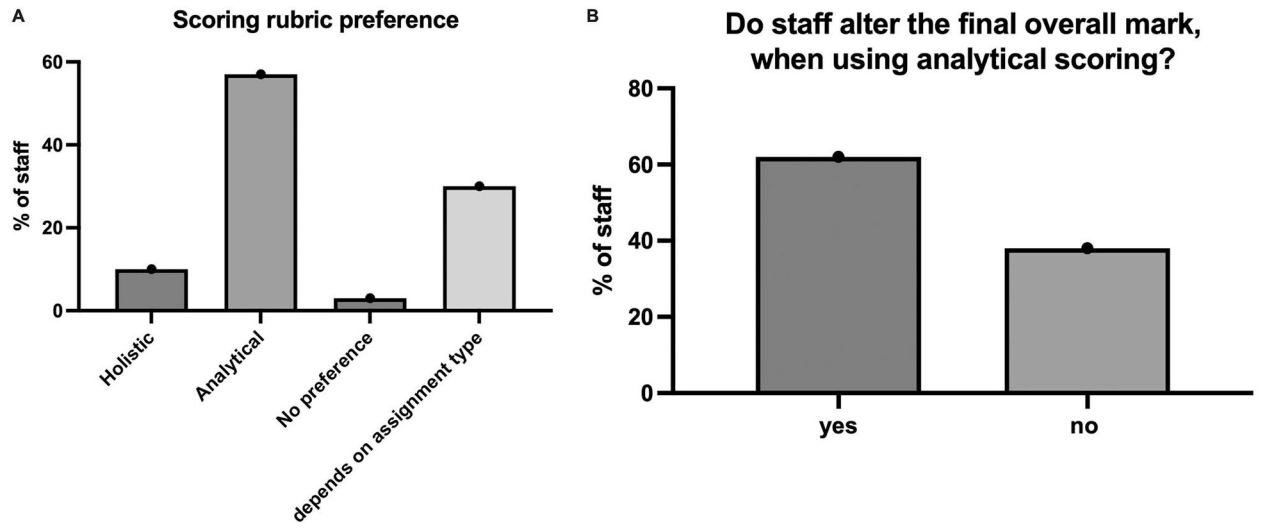

FIGURE 5
The majority of staff that prefer to use an analytic scoring rubric **(A)**, with several having a different preference depending on the assignment being scored. Many staff alter the calculated mark if scoring analytically **(B)**.

However, respondents using holistic scoring have various ways of avoiding bias with one participant stating: "*I tend to go back & remark earlier work to ensure my grades are consistent*".

3.4.3 Theme 3: system constraints and "rule-breaking"

A dominant finding was the tendency for scorers to over-rule calculated grades (42 mentions). While participants prefer to mark analytically, 63% of these scorers alter the final mark if it disagrees with their holistic impression (Figure 5). While this can be due to a belief that holistic scoring is more consistent: "*holistic marking is done 'unofficially' in the background to ensure overall marks are sensible and consistent with marking across the cohort*", this "rule-breaking" is often a reaction to an "inflexible" rubric: "*I find the analytic method is not nuanced enough to allow for all of the work that the student has done to be assessed and it is difficult when work partially falls into several categories*". However, participants also suggested a lack of trust in the validity of the grade: "*using analytical components can sometimes result in a grade that—to me—does not always match what I informally or subjectively feel is the holistic quality and overall category of grade*".

When an assigned mark falls on a grade boundary, this also can result in inflation or deflation of the grade: "*Marks which fall just below grade boundaries are tricky and may be contested by students—in those instances I would consider whether I felt the higher grade boundary was justified for the work overall*" and the converse: "*When section marks lie on the grade boundaries I may mark down if I believe overall the submitted assessment is below the grade boundary*".

Few responders claimed to never alter a grade, with a "sense of correctness" often being the cause: "*if I altered the marks I would not be following the criteria—which is the point of making a criteria in the first place*". Interestingly, the majority of scorers that claim to never alter an analytically determined grade were inexperienced or relatively inexperienced scorers (64%).

3.4.4 Theme 4: challenges to evaluative clarity

The preference for scoring rubrics altered based on the type of assignment and scorer experience (Figure 5). Experienced scorers showed a stronger preference for analytical scoring when unfamiliar with content, yet they were the most likely to use those marks only as a guide, noting that "*analytical marks help you to come to a holistic conclusion.*"

Conversely, inexperienced scorers relied on the analytical rubric as a rigid "safety net" and are more likely to rigidly stay with the grade awarded by the defined weightings. The concerns inexperienced scorers have in using holistic scoring was acknowledged by experienced staff, who noted: "*The hardest issue is then training new markers as it can take a little while for them to develop judgements needed to use a holistic scheme.*"

4 Discussion

Despite many publications describing the use of rubrics in assessing work, few of these are higher education specific and even fewer assess the judgement ability of the scorers (Brookhart, 2018; Young, 2013). Therefore, this study aimed to address this gap and determine whether holistic or analytic scoring gave less variation in grades for a University-level scientific coursework essay, where the scorers had varied experience of marking work.

The study found that individual scorers are internally consistent, with both holistic and analytic scoring rubrics yielding comparable scores awarded for individual essays by the same scorer, regardless of whether the scorer is experienced or inexperienced. This is true for both holistic and analytic scale descriptors. These data cannot have resulted from a halo effect,

with the scorers altering the analytic score to match the holistic one, or vice versa, as the scorers were blind to the weighting assigned to each category of the analytic scoring and were basing their score solely on the scale descriptor information.

However, although intra-scorer reliability is high throughout, analysis of the agreement between different scorers reveals that reliability is significantly influenced by the type of scale descriptor used. When comparing scores for individual essays, agreement between experienced and inexperienced scorers is poor when using a holistic scale descriptor. This improves to a moderate level of agreement when using an analytical scale descriptor, but only when analytical scoring is used. A similar pattern was seen among experienced scorers; agreement on individual student performance was poor using a holistic scale descriptor but increased to a moderate level when using an analytical scale descriptor, regardless of the scoring method. This suggests that the granular detail of analytical scale descriptors provides a necessary shared framework that reduces inter-scorer variability, although isn't the sole solution to achieving a high level of agreement.

A discrepancy in the actual marks awarded is seen between different scorer groups, regardless of scale descriptor or scoring rubric used. Separate two-way LMM analysis of scores awarded using each scale descriptor showed experienced scorers award significantly lower marks than inexperienced markers for both holistic and analytical scoring when using an analytical scale descriptor, with no difference seen with a holistic scale descriptor. This suggests that while the increased complexity of analytical scale descriptors allows for a better alignment on individual student rankings, it may cause latent differences in professional judgement, with experienced scorers applying more stringent internal standards to criteria than inexperienced scorers. These differences in stringency are not seen when scorers use the broader, holistic scale descriptors. Importantly, the significant covariate effect of the real mark in this analysis confirms that scorers' judgment was anchored in the students' actual performance rather than being fundamentally altered by the rubric type or scale descriptor format.

When conducting a more comprehensive three-way LMM, evaluating the overarching relationship between scale descriptor type, scorer experience and scoring rubric a significant two-way interaction between experience and scale descriptor is seen. This confirms that the impact of scorer expertise isn't uniform, but that the divergence between experienced and inexperienced scorers is significantly influenced by the type of scale descriptor used. This interaction is supported by the estimated marginal means (adjusted for the real mark covariate) which shows experienced scorers maintain a high level of stability across all conditions whereas inexperienced scorers are sensitive to the granularity of analytical scale descriptors. This suggests that while experienced judgment remains consistent, detailed criteria may trigger inexperienced scorers to award more lenient marks by adopting a 'checklist' marking style (Bloxham and Boyd, 2007).

In contrast, the comprehensive three-way LMM showed all interactions involving the scoring rubric were non-significant, suggesting that the structural process of marking doesn't introduce further bias into the results. The real mark also remained a highly significant predictor of awarded scores, indicating that, despite variations in scoring tool or experience, scorer judgment remained primarily anchored in the students' actual performance.

Investigation of the underlying causes of the variance seen between scores, by analysing the reliability of individual sub-scores within the analytical rubric, showed that while scorers may agree on the "real mark," they often lack consensus on the specific categorical breakdown of a grade. When using a holistic scale descriptor, correlations were significant but poor across nearly all categories. This reliability didn't improve with the use of an analytical scale descriptor; in fact, several correlations became non-significant. This suggests that the more detailed scale descriptors failed to provide the intended shared objective framework for judgment. Furthermore, the MAE remained consistently high and didn't decrease as scorers transitioned from holistic to analytical tools, indicating that increasing the granularity of a rubric does not effectively reduce the divergence between individual scorers. Collectively, these findings suggest that while analytical tools may offer the appearance of precision, the increased complexity does not necessarily translate into a more unified interpretation of specific marking criteria.

These data, however, only give an insight into inter-scorer reliability and don't give an indication of the actual accuracy of scores awarded by each experience group. Interestingly, when comparing the scores awarded by scorers within this study with the real scores the coursework essays were awarded, experienced scorers were consistently more likely to award a score close to the real mark than inexperienced scorers. This high level of accuracy among experienced scorers remained stable regardless of the scoring rubric or scale descriptor used. When comparing final grade classification, rather than absolute scores, this pattern was repeated, although, for inexperienced scorers, a holistic scale descriptor was slightly more likely to produce a grade class that agreed with the real essay classification, than an analytic one. This suggests scorer experience is more important in determining accuracy of marking than the type of scoring rubric and scale descriptor used.

Taken together, these findings suggest that neither the scale descriptor nor the scoring rubric functions as the primary driver of the final mark. Instead, variance in awarded scores appears to be rooted in scorer experience and idiosyncratic differences in professional judgment. While assessment tools provide a necessary framework for evaluation, they do not supersede the internal standards and 'professional vision' that markers bring to the task.

Previous research on which scoring method yields higher reliability and validity is mixed and inconclusive (Barkaoui, 2011; Brookhart, 2018; Jönsson and Svingby, 2007; Reddy and Andrade, 2010). A higher inter-rater reliability with holistic scoring is reported by some, whereas others suggest analytic rubrics can be more reliable because they focus raters on specific criteria (Barkaoui, 2011; Çetin, 2013) and that high agreement on holistic scores can sometimes mask scorer divergence on specific criteria, raising concerns about validity (Harsch and Martin, 2013). The finding of this study that neither holistic nor analytic scoring is preferable to the other would appear to be at odds with this research, although differences in study design could account for many of the differences reported (Brookhart, 2018; Tomas et al., 2019). The lack of consistent definitions for 'holistic' and 'analytic' rubrics

across different studies also makes it challenging to draw definitive conclusions from comparative research (Dawson, 2017). What one study considers 'holistic multi-trait' might be labelled 'analytic' by another. This inconsistency can lead to seemingly contradictory findings (Brookhart, 2018). The effectiveness of each approach also appears to be highly context-dependent and influenced by factors such as the assessment purpose, subject being measured, and the specific design of the rubric. Additionally, some studies have acknowledged limitations (e.g., small sample sizes) that could affect the conclusions drawn.

Our study shows inter-rater reliability is similar when raters score the same essay, regardless of scoring rubric type. This appears to contradict the findings of a study that performed a similar analysis and reported that the strongest correlations occurred between two holistic raters followed by two analytic raters (Çetin, 2013). The study also reported that inter-rater reliability is low when holistic marks are compared to analytic for the same student essay. However, this study compared marks awarded exclusively by inexperienced scorers. Thus, the findings cannot be directly compared as our study generated data from pairings involving at least one experienced scorer (either mixed-experience or two experienced scorers) for every essay. Our findings are more in agreement with an earlier study, however, which compared experienced scorers to each other, and found a strong correlation (Wang, 2009), although also reported that analytic scores tended to be higher than holistic ones overall (which wasn't found in our study).

One further finding of our study is that individual scorers are internally consistent (high intra-scorer reliability) but lack inter-scorer alignment. This suggests that the "score gap" isn't a result of random error but is a systematic difference between how two independent scorers interpret the same criteria. This gap is then widened if the scorers differ in experience. This conflicts with a previous report that found experienced scorers don't show a significant difference in scores when compared to each other (Lim, 2011) and that inexperienced scorers quickly adapt to the rating-level of an experienced counterpart. However, this study used a holistic 10-point scoring scale, which could lend itself to less variation than the 100-point scale used within this study. Scorers are much more consistent when choosing between 5 and 9 categories than when choosing between 100 individual points, due to the influence of human cognitive limits (Yorke, 2011), suggesting that the 100-point scale in our study provides more opportunity for individual variance to be seen.

It's important to note that scorers in this study were not given formal training to help calibrate their scoring prior to assessing the essays. While the literature widely advocates for scorers to be trained, to enhance inter-rater reliability (e.g., Attali, 2015, Eckes, 2008 Hodges et al., 2019;; Jönsson and Svingby, 2007), training was intentionally omitted in this study to prevent it from functioning as a confounding variable. By excluding this calibration phase, the study was able to isolate the 'raw' impact of the scale descriptors on scorers of varying expertise, reflecting a potentially common use of these tools in high-pressure Higher Education environments where extensive training is not always feasible. The data from this study consequently allows for future assessments to be focused on the most inherently effective scale descriptor and scoring rubric, with the understanding that training can then be provided to further enhance their efficacy. Furthermore, while raw scores between experienced scorers in our study showed some variation, their independent agreement with the real moderated essay grade was strong. This suggests that, for experienced scorers, internalised standards may provide a high degree of evaluative accuracy (Bloxham et al., 2011) that reduces the need for detailed formal calibration. This finding refines the common academic consensus on the essentiality of rater training, particularly for experienced staff, suggesting different training may be delivered depending on experience level.

Despite the data from this study showing no difference in the accuracy of grades awarded between analytic and holistic scoring, it is clear that most staff prefer to mark analytically, although the majority of these then admit to altering the overall score if it doesn't agree with the holistic grade they formulate during the assessment process. This suggests that analytical scoring is often used as a guide to confirm a holistically determined mark, which agrees with previous reports of lecturers making holistic judgements when verbalising their thinking as they scored, with many not making use of written criteria during scoring and only referring to it to justify their holistic decision (Bloxham et al., 2011). This is perhaps also supported by the finding that staff change their preference for scoring rubric depending on the familiarity of the assignment they are marking. The less familiar they are in marking the assignment type, the stronger their preference for analytic scoring, which calculates the overall score for the scorer. This is replicated in the preference of inexperienced scorers, who are also more likely to prefer analytical scoring and are more likely to rigidly stay with the grade awarded by the defined weightings, often citing a lack of confidence or knowledge as the barrier to deciding a holistic grade. Several scorers expressed a belief that analytical scoring leads to less variance between markers, which the quantitative part of this study has proven to not be the case. While some scorers prefer the flexibility of holistic scoring to decide a grade based on the overall criteria, considering aspects the marking criteria isn't explicit on, or that falls between grade boundaries, others are concerned about fairness and bias. This is mostly a fear of a grade being influenced by a significant strength or weakness in one area attracting the focus of the scorer. However, this study revealed that scorers also have various ways of avoiding bias, such as marking all assignments in a single day, re-evaluating the cohort once all are graded, and by double marking (rather than moderation). These actions may explain the lack of difference in accuracy of grading between the two scoring rubrics seen in this study. The belief in the reliability of analytical scoring found in this study contradicts previous research, however, where critics argue that such rubrics lack precision and only provide an illusion of accuracy (Kohn, 2006; Panadero and Jonsson, 2020; Royce Sadler, 2009; Wilson, 2007). These critics reason that the list of criteria can never be complete and some characteristics being evaluated are impossible to explain within a rubric.

The fact that the opinions of staff in this study regarding scoring rubric reliability differ from the empirical evidence agrees with previously reported findings that many scorers who are critical of a rubric type refer to anecdotal evidence or personal experience, while research that generates actual data often refutes these claims (Panadero and Jonsson, 2020; Royce Sadler, 2009). This highlights the importance for basing

decisions on rubric use on scientific data rather than anecdotal evidence. This study demonstrates that both holistic and analytical scoring rubrics yield reliable results when the scorers are experienced, supporting previous findings (Brookhart and Chen, 2015). Thus, it may be advisable to undertake a scorer validation exercise, incorporating training for inexperienced scorers, prior to assessing coursework to avoid misplaced scorer preference from affecting their assessment.

The finding that the type of scoring rubric and guidance has no effect on grades awarded also suggests that the rubric used can be chosen based more on how students engage with these rubrics than with staff engagement. If students find analytic more easily understood than holistic (or vice versa) with respect to matching their grade to the reasoning behind how it was awarded, then this could increase feedback literacy and encourage greater engagement with feedback to improve learning (Carless and Boud, 2018; Malecka et al., 2022).

This study does have limitations. The definition of an inexperienced scorer is subjective and based on familiarity with assessing a final year undergraduate coursework essay with less consideration taken over the knowledge of the science contained within the essay. The scorers in this study all had subject knowledge, but were not all equally experienced in assessing coursework and the specific criteria this involves, some of which are subjective (e.g., synthesis of writing, use of references, essay structure). This introduces a limitation in that the scorers taking part in this study were potentially gaining experience over the four-year study period. To account for potential scorer maturation, we conducted a sensitivity analysis comparing early-stage and late-stage accuracy within the inexperienced scorer group and found no significant longitudinal change, suggesting no systematic improvement in scorer proficiency. The study also included a reference rater (anchor) who assessed the entire cohort of essays and served as a longitudinal control. The anchor's standards were verified for stability across all four years, providing a constant baseline against which to measure any potential scorer instability in the inexperienced pool. Our results show that the relative distance between inexperienced marks and the anchor remained consistent throughout the study, which strengthens the suggestion that inexperienced scorers were not gaining enough experience within the study duration to introduce a confounder. A further limitation is that the scores are generated from the marking of coursework essays that involve assessing a range of criteria, some of which are subjective. It is possible that other assignments, that are scored solely on factual content will have a higher agreement between scorers and be less dependent on the experience of the scorer.

Sample sizes in some of the analyses conducted within this study are also limitations. In the quantitative, empirical study, the second phase assessing scores awarded using an analytical scale descriptor has a relatively modest sample size of 62 essays. While this sample provided sufficient data for an initial comparison of holistic vs. analytical scale descriptors, the lower statistical power compared to the longitudinal first phase (using a holistic scale descriptor) means the findings should be interpreted as indicative rather than definitive (Bolker et al., 2009). Larger-scale replication is required to confirm the extent to which these specific descriptors affect scorer reliability and accuracy across diverse essay types (Westfall et al., 2014).

Further to this is the relatively small sample size of the qualitative survey, which consisted of 30 respondents (a 32% response rate). While this exceeds typical benchmarks for internal staff surveys (Baruch and Holtom, 2008) and allowed for the achievement of thematic saturation (Guest et al., 2006), the findings should be viewed as an exploratory insight into staff perceptions within this specific institutional context rather than a generalisable consensus.

The qualitative sample also consisted primarily of staff who self-identified as 'experienced', accounting for 86% of the total responses. While the inclusion of 'relatively experienced' and 'inexperienced' scorers provided an important counter-perspective, the findings primarily reflect the views of staff with high institutional literacy. Future research could specifically target new faculty members to determine if the analytical descriptors provide a different level of 'scaffolding' for those without extensive prior marking experience.

Furthermore, the qualitative data were processed by a single coder. Although a systematic thematic framework and peer debriefing were employed to ensure the dependability of the results, the absence of independent inter-coder agreement introduces a potential for subjective bias in the interpretation of the themes. Future studies could address these constraints by using multiple independent coders to determine inter-coder reliability and provide a more robust interpretation of the data.

A primary strength of this study was the methodological requirement for scorers to provide a holistic score prior to an analytical score, while remaining blinded to the analytical weightings. This sequence captured the scorer's unbiased global impression and ensured that analytical scores remained independent data points rather than components of a known formula. Expert judgment is often an integrated process (Bloxham and Boyd, 2007), thus we prevented scorers from using an internal check list to calculate a grade, forcing them instead to rely on the professional "gut instinct" required to test their judgment against the criteria.

Furthermore, blinding scorers to weightings prevented the manipulation of analytical scores to match an internal pre-determined holistic impression. This allowed for a true metric of internal scorer consistency to be determined, which would reveal a divergence between scores (such as a high analytical score but a low holistic one revealing instances where a scorer "felt" an essay was poor despite the presence of required criteria). Finally, this approach revealed scorers' implicit weightings, enabling statistical modelling to accurately determine which criteria were the strongest predictors of the holistic scores awarded.

By contrast, providing the analytical rubric first would have forced focus onto isolated parts, inhibiting the holistic judgment and risking a reductive, "tick-box" approach (Bloxham and Boyd, 2007). This would result in a calculated reflection causing the productions of a holistic grade that matches the breakdown scores to avoid appearing inconsistent, effectively masking the scorer's true values and the underlying behaviour of their judgement.

In conclusion, regardless of which type of scoring rubric is used, it's clear that most scorers prefer to use analytical scale descriptors that clearly define expectations for individual criteria. This preference may be determined by the use of this

guidance to inform the wording of their feedback rather than the guidance being more informative in deciding the grade awarded, as many scorers determine the grade they award holistically, either by directly holistically scoring or by altering the analytically awarded grade to better fit their idea of an overall score formed during the assessment, as they gain experience in marking an assignment. The empirical findings of this study suggests that the type of scale descriptor and scoring rubric used have no effect on the variance of scores awarded, and therefore staff preference can legitimately be considered when formulating the assessment of assignments. However, perhaps a more important consideration is that of student preference and confidence in the fairness of the assessment procedure, with the scale descriptor and scoring rubric designed to address the student perspective as a priority. Regardless of assessment choice, our findings suggest that scorer experience acts as a robust variable that is relatively immune to changes in scoring rubrics or scale descriptors. The lack of interaction between experience and scoring rubric indicates that simply providing analytical scale descriptors is insufficient to reduce the variation in scores and accuracy between inexperienced and experienced scorers. Inexperienced scorers need time to gain experience before their scores can be considered reliable, and an inexperienced scorer potentially should be moderated or double-marked by an experienced scorer to ensure consistency of grade alignment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: University of Sheffield research data repository. https://doi.org/10.15131/shef.data.30086350.

Ethics statement

University of Sheffield Ethical Review Panel. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. SF: Conceptualization, Writing – review & editing. TH: Validation, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Acknowledgments

The authors acknowledge the use of Google Gemini 3 Flash for assistance in wording the methodological descriptions to ensure technical accuracy.

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 used in the creation of this manuscript. The authors acknowledge the use of Google Gemini 3 Flash for assistance in wording the methodological descriptions to ensure technical accuracy.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2026.1729644/full#supplementary-material

References

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Lang. Test.* 33 (1), 99–115. doi: 10.1177/0265532215582283

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System* 29 (3), 371–383. doi: 10.1016/S0346-251X(01)00025-2

Barkaoui, K. (2010). Variability in ESL essay rating processes: the role of the rating scale and rater experience. *Lang. Assess. Q.* 7, 54–74. doi: 10.1080/15434300903464418

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. Assessment in education: principles. *Policy Pract.* 18 (3), 279–293. doi: 10.1080/0969594X.2010.526585

Baruch, Y., and Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Hum. Relat.* 61 (8), 1139–1160. doi: 10.1177/0018726708094863

Bloxham, S., and Boyd, P. (2007). *Developing Effective Assessment in Higher Education: A Practical Guide*. New York: Open University Press/McGraw-Hill Education.

Bloxham, S., Boyd, P., and Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Stud. High. Educ.* 36 (6), 655–670. doi: 10.1080/03075071003777716

Bloxham, S., den-Outer, B., Hudson, J., and Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assess. Eval. High. Educ.* 41 (3), 466–481. doi: 10.1080/02602938.2015.1024607

Bolker, B., Brookes, M., Clark, C., Geange, S., Poulson, J., Stevens, M., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24 (3), 127–135. doi: 10.1016/j.tree.2008.10.008

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3 (2), 77–101. doi: 10.1191/1478088706qp063oa

Braun, V., and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport. Exercise and Health* 11 (4), 589–597. doi: 10.1080/2159676X.2019.1628806

Brookhart, S. M. (2018). Appropriate criteria: key to effective rubrics. *Front. Educ.* 3, 22. doi: 10.3389/feduc.2018.00022

Brookhart, S. M., and Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educ. Rev.* 67 (3), 343–368. doi: 10.1080/00131911.2014.929565

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., et al. (2016). A century of grading research: meaning and value in the most common educational measure: meaning and value in the most common educational measure. *Rev. Educ. Res.* 86 (4), 803–848. doi: 10.3102/0034654316672069

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12 (1), 1–15. doi: 10.1177/026553229501200101

Carless, D., and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assess. Eval. High. Educ.* 43 (8), 1315–1325. doi: 10.1080/02602938.2018.1463354

Chan, C. K. Y., and Luk, L. Y. Y. (2022). 'Going 'grade-free'?—teachers' and students' perceived value and grading preferences for holistic competency assessment'. *High. Educ. Res. Dev.* 41 (3), 647–664. doi: 10.1080/07294360.2021.1877628

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edn. New York: Lawrence Erlbaum Associates.

Çetin, Y. (2013). Reliability of raters for writing assessment: analytic—holistic, analytic—analytic, holistic—holistic. *Mustafa Kemal Üniv. Sosyal Bilimler Enstitüsü Dergisi* 8 (16), 471–486.

Davis, L. (2018). "Analytic, holistic, and primary trait marking scales," *The TESOL Encyclopedia of English Language Teaching*, eds. J. I. Liontas,and M. DelliCarpini (Hoboken, NJ: Wiley-Blackwell), 1–6. doi: 10.1002/9781118784235.eelt0365

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assess. Eval. High. Educ.* 42 (3), 347–360. doi: 10.1080/02602938.2015.1111294

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Lang. Test.* 25 (2), 155–185. doi: 10.1177/0265532207086780

Fulcher, G. (2003). *Testing Second Language Speaking*. New York, NY: Pearson. doi: 10.4324/9781315837376

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *J. Res. Dev. Educ.* 27 (2), 73–82.

Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough?: an experiment with data saturation and variability: an experiment with data saturation and variability. *Field. Methods* 18 (1), 59–82. doi: 10.1177/1525822X05279903

Hamp-Lyons, L. (1995). Rating nonnative writing: the trouble with holistic scoring. *TESOL Q.* 29 (4), 759–762. doi: 10.2307/3588173

Harsch, C., and Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability, assessment in education: principles. *Policy Pract.* 20 (3), 281–307. doi: 10.1080/0969594X.2012.742422

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., and McTigue, E. (2019). Developing and examining validity evidence for the writing rubric to inform teacher educators (WRITE). *Assess. Writ.* 40, 1–13. doi: 10.1016/j.asw.2019.03.001

Jönsson, A., Balan, A., and Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assess. Educ. Princ. Policy Pract.* 28 (3), 212–227. doi: 10.1080/0969594X.2021.1884041

Jönsson, A., and Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educ. Res. Rev.* 22, 130–144. doi: 10.1016/j.edurev.2007.05.002

Kohn, A. (2006). Speaking my mind: the trouble with rubrics. *Engl. J.* 95 (4), 12–15. doi: 10.2307/30047080

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15 (2), 155–163. doi: 10.1016/j.jcm.2016.02.012

Lai, E. R., Wolfe, E. W., and Vickers, D. (2015). Differentiation of illusory and true halo in writing scores. *Educ Psychol Meas.* 75 (1), 102–125. doi: 10.1177/0013164414530990

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33 (1), 159–174. doi: 10.2307/2529310

Lee, Y., Gentile, C., and Kantor, R. (2009). Toward automated multi-trait scoring of essays: investigating links among holistic, analytic, and text feature scores. *Appl. Linguist.* 31 (3), 39–417. doi: 10.1093/applin/amp040

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: a longitudinal study of new and experienced raters. *Lang. Test.* 28 (4), 543–560. doi: 10.1177/0265532211406422

Lopera-Oquendo, C., Lipnevich, A. A., and Mañez, I. (2024). Rating writing: comparison of holistic and analytic grading approaches in pre-service teachers. *Learn. Instr.* 94, 101992. doi: 10.1016/j.learninstruc.2024.101992

Malecka, B., Boud, D., and Carless, D. (2022). Eliciting, processing and enacting feedback: mechanisms for embedding student feedback literacy within the curriculum. *Teach. High. Educ.* 27 (7), 908–922. doi: 10.1080/13562517.2020.1754784

Meteyard, L., and Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *J. Mem. Lang.* 112, 104092. doi: 10.1016/j.jml.2020.104092

Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., and Reble, R. (2022). Judgment accuracy of German student texts: do teacher experience and content knowledge matter? *Teach. Teach. Educ.* 119, 103879. doi: 10.1016/j.tate.2022.103879

Panadero, E., and Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educ. Res. Rev.* 30, doi: 10.1016/j.edurev.2020.100329

Peat, J., Elliott, E., Baur, L., and Keena, V. (2013). *Scientific Writing: Easy When you Know how*. John Wiley & Sons.

Reddy, Y. M., and Andrade, H. (2010). A review of rubric use in higher education. *Assess. Eval. High. Educ.* 35, 435–488. doi: 10.1080/02602930902862859

Royce Sadler, D. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assess. Eval. High. Educ.* 34 (2), 159–179. doi: 10.1080/02602930801956059

Sheppard, B. (2019). Checking your analytic performance rubrics for a halo effect. *ORTESOL J.* 36, 28–33.

Tomas, C, Whitt, E, Lavelle-Hill, R, and Severn, K. Modeling holistic marks with analytic rubrics. *Front. Educ.* 2019 (4, p. 89). Frontiers Media SA. doi: 10.3389/feduc.2019.00089

Wang, P. (2009). The inter-rater reliability in scoring composition. *Engl. Lang. Teach.* 2 (3), 39–43. doi: 10.5539/elt.v2n3p39

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press. doi: 10.1017/CBO9780511732997

Westfall, J., Kenny, D. A., and Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *J. Exp. Psychol. Gen.* 143 (5), 2020–2045. doi: 10.1037/xge0000014

Wilson, M. (2007). Why I won't be using rubrics to respond to Students' writing. *Engl. J.* 96 (4), 62–66. doi: 10.2307/30047167

Wolfe, E. W., Moulder, B. C., and Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a rasch multi-faceted rating scale model. *J. Appl. Meas.* 2 (3), 256–280.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Lang. Test.* 24 (2), 251–286. doi: 10.1177/0265532207076365

Yorke, M. (2011). Summative assessment: dealing with the 'measurement fallacy'. *Stud. High. Educ.* 36 (3), 251–273. doi: 10.1080/03075070903545082

Young, C. (2013). Initiating self-assessment strategies in novice physiotherapy students: a method case study. *Assess. Eval. High. Educ.* 38, 998–1011. doi: 10.1080/02602938.2013.771255