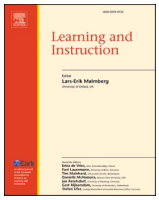
Learning and Instruction 94 (2024) 101992

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Rating writing: Comparison of holistic and analytic grading approaches in pre-service teachers

Carolina Lopera-Oquendo [a,*], Anastasiya A. Lipnevich [b], Ignacio Mañez [c]

[a] The Graduate Center, City University of New York, USA
[b] Queens College and the Graduate Center, City University of New York, USA
[c] ERI-Lectura, University of Valencia, Spain

ARTICLE INFO

Keywords:
Grades
Pre-service teachers
Holistic scoring
Analytic scoring
Rubrics

ABSTRACT

Background: In a typical instructional setting, teachers are responsible for making ongoing decisions that involve judgments of students' capabilities, knowledge, learning needs, and progress toward a certain pre-specified goal. However, there is a significant within-teacher as well as a great between-teacher variability in the actual determination of grades. Grades appear to be an amalgam of characteristics of a student, filtered through a range of teacher personality variables.

Aims: The purpose of this study was to investigate the extent to which pre-service teachers agreed on students' grades in writing task between holistic and analytic grading approaches and how their individual characteristics and beliefs about features of assessment explained the variability in grading practices.

Sample: Teacher candidates (N = 231, 65% female) enrolled in a training program in 2020 and 2021 cohorts at the University of València, Spain, were asked to read two essays, identified by experts as being of low and high quality, and assign holistic and analytic grades.

Results: although teacher candidates provided grades consistently across the two approaches (intra-individual differences), there was a high variability in the distribution among participants (inter-individual differences). We found that, gender, area of specialization, attitudes toward feedback, and extraversion were significant predictors of grading variability.

Conclusion: This study highlights the considerable variation in grading practices among pre-service teachers, indicating the influence of individual factors such as gender, specialization, feedback receptivity, and extraversion. Despite consistent grading within specific approaches, the inter-individual differences in scores were substantial. Due to the consequential nature of teacher grades, our findings offer important insights and have critical implications for teacher preparation and professional development programs.

1. Introduction

In a typical instructional setting, teachers are responsible for making ongoing decisions that involve judgments of students' capabilities, knowledge, learning needs, and progress toward a certain pre-specified goal. In most educational systems, grading decisions represent an idiosyncratic process wherein teachers have to balance their knowledge and beliefs, classroom reality, and external pressures (Brookhart et al., 2016; McMillan et al., 2002; McMillan & Nash, 2000). School grades are a form of feedback that communicates information about students' academic performance and/or progress, which is typically delivered for summative purposes. Hence, grades may have a strong influence on students' sense of achievement, motivation, level of engagement in future courses, and may determine students' educational track and career (Klapp, 2016; Lavy & Sand, 2016; Protivínský & Münich, 2018). So, poor grades assigned by teachers in elementary school are a risk factor for dropout in both middle and high school (Alexander et al., 2001; Bowers & Sprott, 2012; Bowers et al., 2013; Lavy & Sand, 2016), and high school grades are one of the most effective predictors of college admissions and first year college grade point average (GPA), even after controlling for cognitive ability, gender, and SES (Betts & Morell, 1999; Borghans et al., 2016; Camara & Echternacht, 2000; Federičová, 2015).

* Corresponding author. Department of Educational Psychology, The Graduate Center, The City University of New York, 365 5th Avenue, New York, NY, 10016, USA.
E-mail address: cloperaoquendo@graduatecenter.cuny.edu (C. Lopera-Oquendo).

https://doi.org/10.1016/j.learninstruc.2024.101992
Received 1 December 2023; Received in revised form 6 June 2024; Accepted 5 August 2024
Available online 23 August 2024

Considering the aforementioned evidence of the key importance of grades, there is a long history of research into teachers' rationale behind their grade assignments (for a review see Brookhart et al., 2016; Brookhart & Nitko, 2014). Across studies, two findings are remarkably consistent. First, although student achievement should be the main factor determining a student's grade, grades commonly include non-cognitive components, such as students' effort or classroom behavior (McMillan, 2001; McMillan et al., 2002). As a result, this medley of criteria on which grades are based often produce unreliable and uninterpretable grades (Brimi, 2011; Brookhart et al., 2016). Second, there is a significant within-teacher as well as a between-teacher variability in the actual determination of grades. This variability is mainly explained by differences in teachers' knowledge, experiences, expectations, and understanding of the meaning and purposes of grades (Bloxham et al., 2016; Guskey & Link, 2019; Read et al., 2005), and by how teachers weigh cognitive and noncognitive factors in their judgment of students' performance (Jönsson et al., 2021; Martínez et al., 2009). These factors affect reliability and validity of grades and may subsequently lead to distorted decisions about students' academic and professional trajectories, particularly in those educational systems where teachers' grades are considered high-stake criteria for promotion, graduation, and post-secondary admission (Brookhart & Nitko, 2014; Cornwell et al., 2013).

Recent studies have examined the consistency of the agreement among teachers and grading models (Brookhart, 2018; Jönsson & Balan, 2018; Jönsson et al., 2021), the variability due to the use of specific models and strategies for grading (Bloxham et al., 2011, 2016; Randall & Engelhard, 2009, 2010; Tomas et al., 2019), as well as the validity of grading (Brookhart & Chen, 2015; Bonner, 2013; Hodges et al., 2019). Although there are several studies that have explored differences between holistic (i.e., teachers assigning a single score or a letter grade) and analytic (i.e., teachers assigning separate grades per different criteria) scoring approaches (e.g., Jönsson et al., 2021; Tomas et al., 2019), a number of important questions remains unanswered. In this study we attempted to compare the consistency of pre-service teacher grades as they used holistic and analytic approaches to grading and observed inter- and intra-individual differences in grade assignment. We also examined how pre-service teachers' beliefs about feedback and their personal characteristics explained this variability.

1.1. Models for teacher grading

As we mentioned above, the two most common approaches to grading and assessment are holistic and analytic (Jönsson & Balan, 2018; Sadler, 2009). In holistic assessments teachers compile all available evidence about students' proficiency to make an overall qualitative judgment and map it directly onto the grading scale, such as a single score or a letter grade. The competing approach, analytic grading, involves making separate qualitative judgments about different aspects of student performance based on a preset of criteria to report differentiated grades (Guskey & Bailey, 2010). Various grading tools or systems such as rubrics, grading schemes, scoring keys or guides, or criteria sheets are considered analytic grades. For the sake of simplicity and clarity, we will refer to the analytic method as "rubrics" or "analytic approach" interchangeably.

The most apparent advantage of using analytic assessments is that they provide a detailed representation of student performance. Consequently, the analytic model may increase scorer reliability or consistency between grades (such as inter-rater reliability) by preserving the connection of teacher qualitative judgments aligned to explicit and shared criteria. However, there is also a risk that the analytic model may negatively influence the grades' validity in terms of construct underrepresentation (Jönsson & Balan, 2018). Researchers have compared the unidimensional and multidimensional grading models and examined their reliability and validity. The existing body of work has generated inconclusive evidence, and there is no clear rationale for the use of either approach (Brookhart, 2018; Jonsson & Svingby, 2007; Reddy & Andrade, 2010; Sadler, 2009).

Overall, holistic marks have been found to offer reasonable reliability (Jones & Alcock, 2014; Tomas et al., 2019). This approach is less time-consuming, but more subjective, so the validity of holistic grading is generally low (Bouwer et al., 2017). However, their psychometric qualities seem to improve when rubrics and exemplars are used and rater training is conducted (Jonsson & Svingby, 2007). Despite their strength, Sadler (2009) discusses several limitations of holistic judgment methods in educational assessments. First, the author argues that holistic judgments are unreliable, especially when the same teachers grading the same work on successive occasions. Moreover, systematic biases, such as the 'halo' effect, where an assessor's personal knowledge of a student's characteristics may affects their judgment (Sanrey et al., 2021), further complicating the grading process. Additionally, biases related to race, ethnicity, or gender can impact the fairness of grades (Quinn, 2020). Assessors also struggle to maintain cognitive consistency over time, leading to trends of increasing leniency or severity (Pliske & Klein, 2003). Lastly, the variability in grading standards adopted by different assessors, often defended as an academic prerogative, contributes to the inconsistency of holistic grades.

Conversely, the analytic approach is far less efficient than holistic marking (Tomas et al., 2019), and as some researchers have argued that focusing on different aspects of student performance can paint a fragmented picture of students' achievement (Sadler, 2009). Similarly, studies suggest that the content validity of holistic marks would inevitably be low because they represent a mix of idiosyncratic criteria and include a wide range of construct-irrelevant aspects, compromising valid score interpretation (Bloxham et al., 2016; Tomas et al., 2019). Jönsson and Balan (2018) argued that multidimensional grades may provide a more valid picture of student proficiency but are more challenging to interpret. In contrast, Sadler (2009) stated that grades derived from analytic approach also lacked validity because it was challenging to represent all the complexity of qualitative judgments using a simplistic combination of rules.

At the moment, very few studies have directly compared analytic and holistic approaches to grading (Jönsson et al., 2021; Klein et al., 1998), especially using robust quantitative methods. For example, some studies suggested that the analytic grading model reduced the complexity of grading, and thus was preferable to holistic grading in terms of intra-rater consistency (Harsch & Martin, 2013; Jonsson & Svingby, 2007). However, when teachers assigned grades, both approaches revealed low agreement. Rezaei and Lovorn (2010) conducted two experiments to investigate the reliability and validity of grades in the assessment of students' writing with and without a rubric. Findings showed that the range and variability of assigned scores increased significantly after using rubrics. Therefore, results did not provide evidence that using rubrics lessened the variability of assigned scores. Similarly, Jönsson and Balan (2018) randomly assigned teachers to grade the same student assignment using either an analytic or holistic approach. Results showed that the analytic condition yielded higher agreement among assessors than the holistic condition, without differences in how teachers described the quality (either positively or negatively) of students' performance.

The evidence from empirical studies that compare holistic and analytic approaches remains inconsistent, with hardly any studies exploring teacher and student variables and their influence on grading using either holistic or analytic approaches. The only exception are studies that analyzed racial bias in teachers' evaluations using the two grading approaches (Quinn, 2020). Hence, a lot remains unknown about how rater characteristics and rating conditions affect scoring, which is what our study intended to uncover.

1.2. What factors influence teachers' decision-making when assigning grades?

In the context of classroom assessment, teachers' grading decisions are notoriously difficult to gauge because they are highly individualized (Cizek et al., 1995; McMillan & Nash, 2000). They combine a wide range of students' cognitive and non-cognitive factors and depend upon teachers' characteristics, beliefs, personal philosophy of teaching and learning, classroom realities, and a number of external factors, such as parents, administrators, and local grading policies (Cizek et al., 1995; McMillan et al., 2000; Tomlinson, 2001; Brookhart, 2014; Brookhart, 2013; McMillan, 2001, 2003; Bowers, 2011; Lekholm et al., 2008; Lekholm & Cliffordson, 2009; Randall & Engelhard, 2009, 2010).

Brookhart et al. (2016) showed that over the past 20 years most studies on multiple influences on grades included students' noncognitive components, such as effort, work habits, attention, and participation; and other "personal factors" related to students' personality and classroom behavior (Brookhart, 2013; Brookhart, 2014, p. 2016; Mc Millan, 2002; Cizek et al., 1995). Similarly, Kunnath (2017) revealed that when teachers considered ability, grades of low-achieving students became more subjective and less accurate, as teachers looked to increase grades with other factors, such as effort. In contrast, grades of high-ability students tended to be more objective and precise as teachers sought to maximize the weight of cognitive factors. Cross et al. (1999) found that 76% of teachers reported they had inflated grades of low-ability students, 82% considered students' growth or relative improvement in their grading, and 51% reported that class participation affected their grading. Further, Randall et al. (2010) examined teachers' grading decisions of borderline grades (e.g., grades at the border of an A and B, B and C) and found that teachers made decisions based more on overall student ability, behavior, and effort as compared to their objective performance. Researchers have also shown that student motivation and engagement strongly influenced grades (Isnawati & Saukah, 2017), and teachers' perceptions of study habits, rule adherence, attitude, students' personality, and classroom participation carried a significant weight in grading decisions (Cheng & Sun, 2015; Duncan & Noonan; 2007; McMillan, 2001).

Despite the number of studies that examined how teachers' characteristics serve as central predictors in models describing assessment processes or moderators of teachers' judgments and judgment biases (Heitzmann et al., 2019; Loibl et al., 2020), quantitative studies that carefully examined how teacher characteristics, such as gender, content expertise, attitudes, and beliefs, effects on grades rationality and variability are relatively scarce. So, for example, research into teachers' gender differences in grading reveals mixed findings. Some studies have found that the gender grading gap is highest with male teachers (Lavy, 2008; Lindahl, 2016), while other researchers' findings did not show support that teachers favor students of the same gender or from the same foreign background as themselves (Doornkamp et al., 2022; Lindah, 2016). Studies have also revealed that grading bias in primary and secondary schools was associated with variation in teachers' expectations talent and effort (Doornkamp et al., 2022) and teachers' beliefs and gender stereotypes (Protivnskýe et al., 2018; Cornwell et al., 2013; Hanna & Linden, 2012; Hinnerich et al., 2011).

Higher qualifications, which include experience, content knowledge, and academic degrees in the relevant subject, are assumed to significantly influence teacher judgments (Blömeke et al., 2015; Simonton, 2003). Consequently, studies on the rationality of teachers' grades suggest that experience and content knowledge may relate to the accuracy of teachers' decision-making in grading (Jansen et al., 2021; McMillan, 2003; McMillan & Nash, 2000). However, the limited empirical evidence on the impact of teacher qualifications on teacher judgments presents heterogeneous evidence regarding its effects on judgment accuracy and grading rationality. Regarding teacher professional experience, most studies compare the absolute judgments of experienced teachers (in-service teachers) with those of teacher candidates (preservice teachers). The majority of these studies reveal that experienced teachers make stricter and less accurate judgments compared to teacher candidates and expert ratings (Barkaoui, 2010; Lim, 2011; Jansen et al., 2021). However, some studies have shown that experience was not a predictor of teacher judgments (Meadows & Billington, 2010; Zhu & Urhahne, 2015). For instance, Guskey (2019) found statistically significant differences among teachers at different grade levels, but no differences related to teachers' years of experience.

Regarding content knowledge, studies suggest that it does not lead to clear differences in teacher judgments either. For example, Meadows and Billington (2010) compared text ratings and found no differences in judgments based on the level of competence of teachers who had studied the subject in question versus those who had studied other subjects. Jansen et al. (2021) compared the judgments (holistic and analytic grades) of in-service and pre-service teachers to machine scores and expert ratings. They found that experienced teachers' judgments were more negative than those of student teachers, and both were more negative than the judgments made by experts and the machine. Interestingly, these results remained stable even after controlling for content knowledge. Conversely, Moller (2022), using a quasi-experimental study design examined whether teacher experience and content knowledge related to teacher judgments on student students' German texts. Results showed that experienced teachers made stricter and more heterogeneous judgments than teacher students and trained experts. However, relative judgment accuracy (correlation between teachers' and experts' judgments) was higher when teachers majored in German Studies.

Researchers have documented the connections between personality traits and measures of job performance across multiple occupations, including teaching (Barrick & Mount, 1991; Bastian, et al., 2015, 2017; Judge et al., 2013; Kim et al., 2018, 2019; Klassen et al., 2017; Klassen & Tze, 2014). Kim, Jörg, and Klassen (2019) conducted a meta-analysis of 25 studies examining the relationships between the Big Five personality domains and teacher effectiveness and burnout. The authors noted that extraversion was more strongly associated with teacher effectiveness than conscientiousness, which has been consistently reported as the stronger predictor of job performance in multiple occupations (Barrick & Mount, 1991; Judge et al., 2013; Salgado, 1997, 2003). Conversely, agreeableness was not associated with teacher effectiveness. Interestingly, Kim et al. (2018) found that although teacher personality predicted subjective measures of teacher effectiveness, it did not predict objective measures, such as student academic achievement in specific areas (e.g., English or mathematics). Although there is no empirical evidence on how personality affects grading decisions in pre-service teachers, it could be expected that the grades assigned by teachers may be more influenced by personality compared to students' standardized test scores.

Grading is a significant component of feedback that teachers routinely provide. Although studies have found that teachers based their grading practices on their educational philosophies and beliefs (McMillan & Nash, 2000; Tomlinson, 2000) and that teachers' knowledge, skills, beliefs, and attitudes regarding assessment practices influenced their approaches to grading (e.g., Fulmer et al., 2015), there is no evidence about whether attitudes toward instructional feedback could affect teachers' grading-related decisions. Receptivity to instructional feedback is described as cognitive, affective, behavioral, and instrumental attitudes toward feedback one receives. It is possible that instructors with high receptivity would be more skilled at providing grades to their students. In other words, we could expect that individuals who are more willing to engage with feedback may be more tuned into discerning relevant information and thus provide more reliable and accurate judgments. In this study we considered this possibility.

All in all, grades appear to be an amalgam of both achievement-related and noncognitive characteristics of a student, filtered through a number of teacher personality variables. In this study, we will examine the interplay of these characteristics, focusing on teacher characteristics

and differential levels of student performance.

1.3. The current study

The purpose of this study is to investigate the extent to which pre-service teachers agreed on students' grades across two different grading approaches and how their beliefs about features of assessment (receptivity to instructional feedback) and their individual characteristics (gender, content expertise, and personality) explained the variability in grading practices. Specifically, this study attempted to answer the following research questions.

1. Are pre-service teacher scores given to writing task essays consistent (intra-rater reliability) across grading methods (holistic vs. analytic)?
2. Is there a significant variability in scores among pre-service teachers (inter-rater reliability) when using different grading methods (holistic vs. analytic)?
3. Can variability in grading be explained by pre-service teachers' gender, attitudes toward feedback, personality traits, and their content expertise?

2. Method

2.1. Participants

255 master's degree students in a Secondary Teacher Education at Universitat de València, Spain, participated in this study.[1] Even though participants were enrolled in the teacher training program, they did not receive specific training on how to provide feedback when assessing written assignments prior to participating in this study. All participants were informed about the study and provided their consent to participate.

An online instrument was administered in two different course cohorts at the end of 2020 and 2021 academic years ($N = 171$ in 2020, $N = 84$ in 2021). Participants who omitted more than 80% of items in the instrument were deleted from the final dataset, so 231 observations were retained ($N = 150$ in 2020, $N = 81$ in 2021). 65% of participants were female, and 76% of the participants were in the age range between 21 and 25 years old ($M = 25.16$, $SD = 4.70$). For the 2020 sample, participants came from different disciplines, 16% were enrolled in the English specialization program, 26% in Language studies and Classical Cultures (Spanish, French, German, Greek, and Latin), 33% in Sciences (Mathematics, Physics, and Chemistry) and 25% in the Social Sciences (Geography and History). The 2021 sample comprised students who specialized in English (47%) and Language studies and Classical Cultures (53%).

2.2. Procedure

The study was conducted through the administration of an online questionnaire designed in Qualtrics. The grading task consisted of providing feedback on two anonymous essays with different levels of performance (one strong and one weak exemplar), then assigning grades using both the analytic rubric and the holistic scoring approach. Later, participants completed the Receptivity to Instructional Feedback scale, the Big Five Personality Inventory and reported their background information. For the purposes of the current paper, we will not consider participants' comments provision and will focus exclusively on their grading of the two essays. The study protocol (2022-0460-QC) was approved by the university's ethics committee.

2.2.1. The grading task

Participants were shown two essays of different levels of student performance. Essays were exemplars from a national standardized test in writing communication in Colombia (SABER T&T) and were classified, according to assessment framework as strong and weak examples. This test assesses learners' ability to communicate ideas through writing on a topic that does not require any specialized knowledge.

The participants were shown the essay prompt and essays written by an anonymous high school student. They were invited to evaluate students' performance level (1 = poor, 2 = fair, 3 = excellent, and 4 = outstanding) in each of the components of the rubric (content, organization, and style) and assign an overall grade based on the quality of student writing in the same way they would in their own classes. Participants received the rubric with the description of performance levels by component (see Table A1, Supplementary Materials). Each participant performed this task twice, once for the strong essay and once for the weak essay. No specific training in how to use this rubric was provided. The sequence of the tasks was not counterbalanced in the first round of data collection (2020); therefore, all participants were exposed first to the strong essay followed by the weak essay. Additionally, participants were asked to provide holistic scores first, followed by analytic scores using the respective rubric. For the second cohort (2021), the type of essay (strong or weak) shown to participants was randomized. However, the sequence of grading methods remained similar to the first cohort, with holistic scores being provided before analytic scores.

2.3. Measures

2.3.1. Scores

Holistic scores were measured as a continuous value ranging from 0 to 10, which is the common scale for providing grades in Spain. Analytic rubric scores (ranging from 1 = poor to 4 = outstanding) for two essays were estimated using the Generalized Partial Credit Model (GPCM; Muraki, 1992). The GPCM, an item response theory (IRT) model, is designed for situations where item responses are organized in two or more ordered categories. In this model, items are conceptualized as a series of ordered steps, with examinees receiving partial credit for successfully completing each step. These steps correspond to various levels of performance required to complete an item. The GPCM is formulated on the assumption that the probability of choosing the *kth* category over the *k*-1-th category is governed by a dichotomous response model. The GPCM to estimate the probability of responding to a specific response category directly (θ) is written as:

$$P(\theta) = \frac{\exp\left[\sum_{k=0}^{K} Da_j\left(\theta - \left(b_j - \delta_{jk}\right)\right)\right]}{\sum_{k=0}^{K} \exp\left[\sum_{k=0}^{K} Da_j\left(\theta - \left(b_j - \delta_{jk}\right)\right)\right]}$$

where *k* is a specific response category in the vector of 0, 1, 2, …K, D is a scaling constant set to 1.7 to approximate the normal ogive model, *a*j is the slope or discrimination parameter of item *j*, b_j denotes the difficulty of item *j*, and δ_{kj} represents the location parameter for a category *kth* on item *j* (Muraki, E., 1997). The slope or discrimination parameter evaluates the measurement quality of an item, indicating the degree to which categorical responses vary among items as θ level changes.

GPCM was selected as method for estimation of analytic scores because this approach is specifically designed to handle items with ordered response categories. It aligns perfectly with the multi-level nature of rubric scoring (e.g., poor, fair, good, excellent). This model allows for partial credit scoring, providing a finer-grained analysis of student performance by awarding partial credit for partial understanding or performance, thus offering more informative feedback than binary

[1] The sample size was determined through power analysis conducted in pwrss R package using probability specification. Assuming a squared multiple correlation of 0.15 between covariables (R-square), a base probability $P_0 = 0.4$. which is the overall probability of being in group 1 without influence of predictors in the model (null), a $P_1 = 0.3$, which is the probability of being in group 1 (P_{11}) deviate from P_0 depending on the value of the predictor under alternative hypothesis, a power of 0.90, and a significance level of 0.05, the recommended sample size for a logistic regression is 216 participants.

scoring methods. Moreover, GPCM enhances the precision of ability estimates by utilizing the full range of rubric scores, leading to more accurate and reliable measurements of student abilities or traits. The model also estimates item difficulty to each step or category in the rubric, the discrimination parameter, and provides detailed item-level statistics for assessing the fit and quality of each rubric item, facilitating both the detailed comprehension of the model parameters associated with each performance level and the identification of poorly performing criteria that may require revision. Overall, GPCM offers a more comprehensive and precise analysis of rubric-based scores in comparison to composite scores, which contributes to enhanced validity and reliability of the assessment (Jabrayilov et al., 2016).

2.3.2. Receptivity to instructional feedback (RIF)

The Receptivity to Instructional Feedback (RIF) scale is a self-report instrument designed to measure individuals' attitudes toward instructional feedback. A total of 24 Likert-type items measured on a 5-point scale (1 = strongly disagree and 5 = strongly agree) were included into three receptivity components: (1) experiential attitudes towards feedback (2) instrumental attitudes; and (3) cognitive engagement with feedback.[2]

2.3.3. Big Five Personality Inventory (BFI)

The Spanish BFI is a 44-item inventory that measures extraversion, agreeableness, conscientiousness, neuroticism, and openness (Benet & John, 1998; Goldberg, 1993). Responses to each personality indicator ranged from 1 (strongly disagree) to 5 (strongly agree).

2.3.4. Demographic and academic background

Participants provided additional information about gender, age, and content expertise, operationalized as their area of specialization (language, mathematics, social science or science).

2.4. Analytic plan

Initially, psychometric analysis to calculate scores for receptivity of instructional feedback scales and analytic scores were estimated, followed by descriptive analyses. For the first research question, Cronbach's alpha was used to estimate raters' consistency among rubric components in analytic scores. The estimation of the agreement among participants' scores and the type of grading model (holistic vs. analytic scores) was assessed using Intraclass Correlation Coefficient (ICC). Spearman's correlation was calculated to analyze the consistency among holistic and analytic grades with scores in each component in the rubric and agreement among rubric's component scores.

For our second research questions, we calculated the consistency among scores provided to each component in the rubric (content, organization, and style) using Fleiss' Kappa, a measure of absolute agreement. Further, we estimated the extent to which participants' scores took similar values to an expert judgment, or consensus agreement (Jonsson & Svingby, 2007; Stemler, 2004). In the current study, consensus agreement was defined as an indicator that assumes a value of 1 when teachers' grades fall within a range of acceptable grades defined by researchers' criteria, and 0 otherwise. These values were determined based on the description of task quality provided by the assessment framework. For strong essays, the range of acceptable values for holistic and analytic scores were higher than 7.5 points or 0.7, respectively. For weak essays, the acceptable grades ranged between 4 and 6 points and lower than −0.5 for holistic and analytic grading approaches, respectively. Additionally, a logistic regression model was estimated to identify individual characteristics that explain the consensus agreement among participants when using holistic or analytic grading models.

Finally, a Hierarchical Linear Modeling (HLM) (Raudenbush & Bryk, 2002) was used for examining variables that can explain grading variability. This approach considers a hierarchical structure of the data, where grades (level-1) provided during the study are nested within pre-service teachers' participants (level-2). This two-level design allows us to calculate the variance component and estimate the effects of both scores (essays' quality and grading method) and the participant's characteristics (gender, specialization area, personality, and attitudes toward feedback) on grading scores. In total, two model were tested. First, we estimated the empty or null model (model 0) that did not contain any explanatory variable and where grading scores (standardized scores) (Yij) were predicted from just an intercept (γ_{00}) and two random effects at level-1 (R_{ij}) and level 2 (U_{0j}). In the subsequent model, grading score features (model 1) and pre-service teacher variables (model 2) were added as a random slope effects and explanatory variables, respectively. A likelihood ratio test on random effects of linear mixed effects model, with χ^2 (chi-square) statistics and corresponding p values, was used to investigate whether each model fit the data better than the previous model.

3. Results

3.1. Psychometric analysis

Psychometrical analysis included the estimation of internal consistency of personality and receptive of feedback scales and the estimation of IRT models for estimation of individual scores of receptivity of instructional feedback scales and holistic grades. Individual scores for analytic grades in the weak and strong essays were calculated using a GPCM. Individual scores were derived from model estimation with a mean of 0 and a standard deviation of 1. Table A2 (Supplementary Material) presents the item parameters and overall item fit test (S-χ^2 tests), while Figs. A1 and A2 (Supplementary Material), displays the Function Information Curve and Item Characteristic Curves, respectively. Overall, the overall difficulty of items for holistic grades ranged from −0.040 (Content) to 0.189 (Organization), while difficulty threshold for each category of response covered a wide range of latent traits.[3] Regarding a-parameter (discrimination) took values from 1.259 (Style) to Organization (3.368), suggesting a high-capacity items to distinguish among different latent traits. Results also indicated a good fit of individual item with **S-**χ^2 p-value $\geq$.01; while RMSEA-χ^2 is a lower cut-off criterion in all cases (p-value < 0.05).

Composite scores for personality traits were derived by adding the answers corresponding to each of the five personality factors. The internal consistency reliability statistics across the five scales ranged from $0.698 < \alpha < 0.864$ (see Table A3 Supplementary Materials). Regarding receptivity of instructional feedback scales, a Confirmatory Factor Analysis (CFA) was conducted to examine the factorial structure of the RIF (see Table A4 Supplementary Materials). Latent variables were derived from an Item Response Theory (IRT) scaling methodology. The Graded Response Model (GRM) (Samejima, 1969), which is an approach for the nominal and polytomous ordinal nature of the items (e.g., rating scales) (Kolen & Brennan, 2014, p. 566); was used to estimate individual parameters and threshold parameters for the items according to the number of response categories (see Table A5 Supplementary Materials) (Bean & Bowen, 2021). Individual scores derived from model estimation were transformed to have a mean of 0 and a standard deviation of 1. The internal consistency reliability statistics (Cronbach's α) across the three scales ranged from $0.774 < \alpha < 0.803$ (see Table A5 Supplementary Materials).

[2] Technical manual containing all the syntax, data, and additional information for scoring the complete scales can be retrieved at https://osf.io/2j6ah/

[3] Range of b-parameter covers the following ranges by scales: *Item 1. Content* from −1.983 to 1.795, *Item 2. Organization* from −1.129 to 1.582, and *Item 3. Style* from −1.865 to 2.174.

3.2. Descriptive information

Fig. 1 presents the distribution of scores for rubric components by the essay quality (weak and strong) in the full sample. The comparative analysis of the "strong essay" and "weak essay" based on the rubric components of style, organization, and content reveals significant differences in performance levels. The strong essay exhibits a higher concentration of *remarkable* and *outstanding* ratings across all three components, with 62% in style, 65% in organization, and 69% in content, respectively. Conversely, the weak essay shows a predominance of *fair* and *poor* ratings, particularly notable in the content component, where 74% of the evaluations fall within these lower categories. Additionally, the weak essay's style and organization components display a considerable proportion of *poor* ratings, 16% and 30% respectively, indicating fundamental deficiencies in clarity and structure. Figs. A3 and A4 (Supplementary Material) presents results by cohort.

Analytic scores, calculated as a latent variable using GPCM model, ranged from −1.510 to 1.836 for the strong essay ($M = 0.394$, $SD = 0.678$) and −2.032 to 2.166 for the weak essay ($M = -0.403$, $SD = 0.905$). Furthermore, holistic grade scores for the strong essay ranged from 2.00 to 9.60 ($M = 7.02$, $SD = 1.28$), whereas the weak essay scores varied from 2.00 to 9.70 ($M = 5.68$, $SD = 1.55$). Descriptive statistics for raw and standardized scores are presented in Table 1. Additionally, Table A6 (Supplementary Materials) presents the descriptive statistics by cohort.

The distribution of standardized grading scores (Fig. 2) demonstrates significant variability, irrespective of essay quality and grading method. This pattern is consistent across all cohorts, as shown in Figs. A3 and A4 in the Supplementary Materials. For holistic grading, scores for strong and weak essays in the 2020 cohort ranged from 2.0 to 9.7 and 1.1 to 9.5, respectively. Similarly, in the 2021 cohort, scores ranged from 2.0 to 9.7 for strong essays and 1.1 to 9.5 for weak essays. In contrast, analytic scores displayed a wider range, spanning from −2.810 to 2.128 in the 2020 cohort and from −2.023 to 2.839 in the 2021 cohort.

Table 1
Descriptive Statistics of Raw and Standardized grades by Essay Quality and Grading Approach.

	Holistic – Strong Essay	Holistic – Weak Essay	Analytic – Strong Essay	Analytic – Weak Essay
Raw scores (N=231)				
Mean	7.018	5.684	0.394	−0.403
Median	7.000	5.500	0.465	−0.546
Std. Deviation	1.279	1.545	0.678	0.905
Minimum	2.000	2.000	−1.510	−2.032
Maximum	9.600	9.700	1.836	2.166
Skewness	−0.491	0.294	0.042	0.645
Kurtosis	0.395	−0.125	−0.212	0.102
Standardized scores				
Mean	0.000	0.000	0.000	0.000
Median	−0.014	−0.119	0.104	−0.158
Std. Deviation	1.000	1.000	1.000	1.000
Minimum	−3.925	−2.385	−2.810	−1.799
Maximum	2.019	2.600	2.128	2.839
Range	5.944	4.985	4.937	4.639

Notes: Standardized scores were calculated for each type of essay and grading approach.

3.3. Intra-rater consistency among participants and grading models

The intra-rater consistency for the analytic scores, measured through Cronbach's alpha, was 0.622 and 0.762 for the strong and weak essays, respectively (see Table A7, Supplementary Materials). The estimation of agreement among participants in standardized analytic and holistic grades is summarized in Table 2. Findings showed a high intra-rater consistency among participants using holistic and analytic grades in strong ($ICC = 0.849$) and weak essays ($ICC = 0.902$). Furthermore, intra-rater consistency among participants by cohort were similar to the results observed in the full sample (see Table A8, Supplementary Materials).

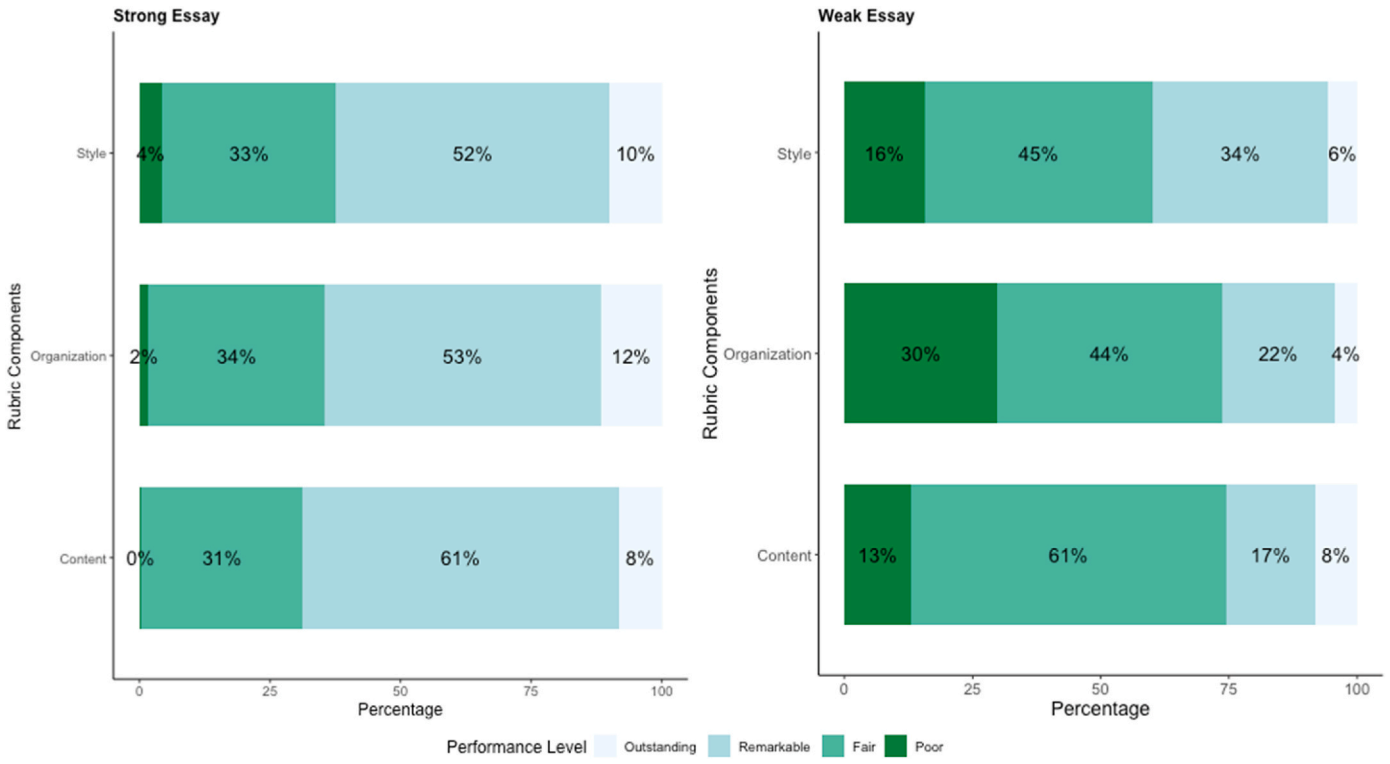


Fig. 1. Distribution of Performance levels in the Rubric Components by Essay Quality and Grading Model for the Full Sample.

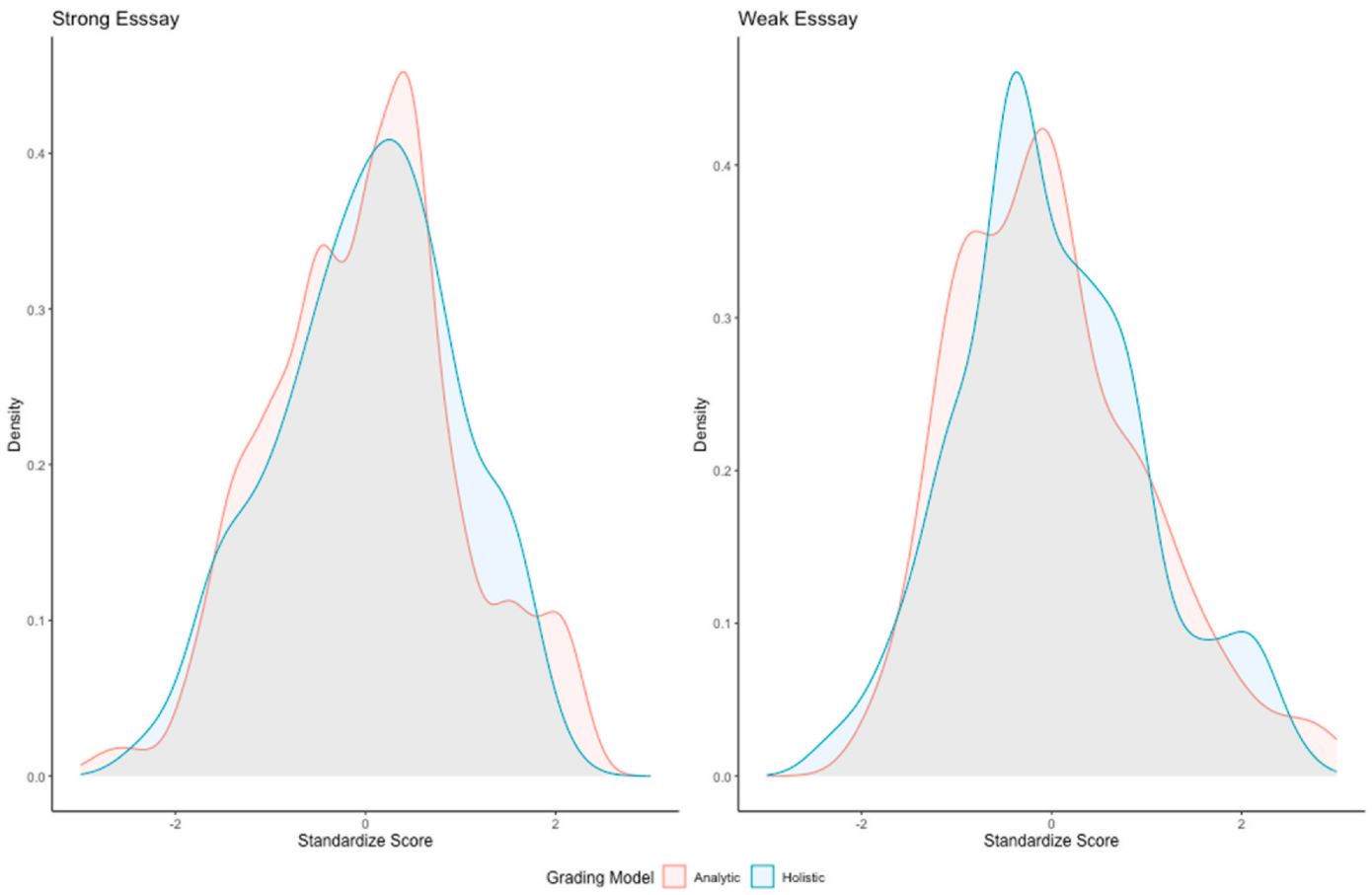


Fig. 2. Distribution of Standardized Scores by Essay Quality and Grading Model. Notes: standantarized scores are included in the plot.

Table 2
Comparison of intra-rater agreement by essay quality.

	ICC	p-value	95% Confidence Interval	
			LL	UL
Full sample (N=231)				
Strong Essay	0.849	<0.001	0.809	0.882
Weak Essay	0.902	<0.001	0.876	0.924

Note: ICC = Intraclass Correlation Coefficient between analytic and holistic grades was calculated using standardized grade scores. Scores were standardized using means and standard deviation for each raw scores in the full sample. LL: Lower Limit, UL: Upper Limit.

Spearman correlation between analytic grades and scores assigned to each rubric component was slightly higher for the weak essay (0.706, 0.930, and 0.730) than the strong essay (0.688, 0.897, and 0.561) (Table 3). Additionally, correlations between the rubric components and scores assigned using a holistic approach suggested a moderate association across all rubric components, which is lower in comparison to analytic scores. The same pattern was observed in both cohorts (Table A9, Supplementary Materials).

3.4. Inter-rater agreement among participants by the type of score and task

Regarding the consistency among the rubric components, the Fleiss' Kappa suggested a slight agreement between participants (0.126 and 0.193 for the strong and weak essays, respectively) (Table 4). These results are also similar in both cohort; however, the agreement between rubric components in 2021 is slightly higher for both weak and strong essays in comparison to 2020 cohort (0.174 vs. 0.142 in strong essay and 0.224 vs. 0.188 in the weak essay). Furthermore, the percentage of agreement among raters was fair (26.84 and 24.68 in the strong and

Table 4
Inter-rater agreement between components of analytic rubric by essay quality.

	Fleiss' Kappa		Percentage of agreement
	Value	*p*-value	
Full Sample (N=231)			
Strong Essay	0.126	<0.001	26.84
Weak Essay	0.193	<0.001	24.68

Table 3
Spearman correlations between rubric components and analytic and holistic scores by essay quality.

	Strong Essay				Weak Essay			
	Analytic score	Holistic score	Content	Organiza-tion	Analytic score	Holistc score	Content	Organiza-tion
Full sample (N=231)								
Content	0.688	0.629			0.706	0.710		
Organization	0.897	0.675	0.411		0.930	0.786	0.501	
Style	0.561	0.658	0.299	0.316	0.730	0.709	0.407	0.572

weak essays, respectively). Similar results were observed by cohort (Table A10, Supplementary Materials).

Regarding the consensus agreement, a higher proportion of participants provided analytic grades consistent with the expert benchmark in comparison to holistic grades in the weak essays (54.1% vs. 46.7%). For the strong essay, the proportion of pre-service teachers who provided scores in agreement with expert criteria was slightly higher in the holistic approach (44.6% vs. 39.8%).

Logistic regression was performed to examine the effects of candidates' gender, cohort, academic specialization, personality, and attitudes toward feedback on the likelihood of pre-service teachers' scores being like those corresponding with the expert criteria. Independent models were run for each essay condition (i.e., weak vs strong) and grading approach (holistic vs analytic). The Hosmer–Lemeshow (H–L) test was used for examining inferences about goodness-of-fit. The H-L test yielded a $\chi^2(8)$ range between 3.193 and 15.43 across all models conducted and was not statistically significant ($p > 0.05$), suggesting that the models provided a good fit the data well. Increases in experiential attitudes toward feedback were associated with an increase in the likelihood of providing holistic ($\beta = 0.505$, $OR = 1.658$, $p = 0.009$) and analytic grades ($\beta = 0.441$, $OR = 1.554$, $p = 0.021$) for the weak essay similar to those provided by experts (see Table A11, Supplementary Materials).

3.5. Explaining variability of grading scores

The first step in the analysis was to explore the components of the variance in pre-service teachers' grades. The within-grade variance ($\sigma^2 = 0.589$) and between-individuals variance ($\tau_{00} = 0.403$) were significantly different from zero. Moreover, results indicated that 40.7% of the total variance in grades (ICC) was attributed to the differences *between* pre-service teachers. An ANOVA showed that random effects were statistically significant ($LRTest = 175.94$, $df = 1$, $p < 0.001$).

Secondly, characteristics of grades (grading approach and essay condition) (model 2) were added as a two random coefficient, to check whether the effect of characteristics of task and grading model affects grades vary across participants. ANOVA suggested that random effects for type of task performance ($LR\ Test = 514.9$, $df = 3$, $p < 0.001$) and grading approach ($LR\ Test = 13.7$, $df = 3$, $p = 0.003$) were statistically significant. The full model added pre-service teachers' background information as explanatory variables (Table 5). Gender (Male = 1, Female = 0) was a statistically significant and negative predictor of grades ($\beta = -0.26$, $p = 0.050$). However, the main effect of gender on grades depended on the cohort ($\beta = 0.10$, $p = 0.006$). Moreover, candidates' extraversion was a statistically significant and positive predictor of grades variability ($\beta = 0.11$, $p = 0.028$). That is, an increase in one standard deviation in extraversion was associated with an increment of 0.11 standard deviation in grades assigned by pre-service teachers. Finally, results also showed that the essay quality ($\tau_{11\ Task} = 1.427$) is the largest source of variability among individuals.

Table 5
Hierarchical linear model estimation.

Predictors	Final Model		
	β	β(SE)	*p*-value
Intercept	0.03	0.10	0.743
Gender [Male]	−0.26*	0.13	0.050
Cohort [2021]	−0.09	0.13	0.477
Gender [Male] * Cohort [2021]	0.70*	0.25	0.006
Area [Science & Math]	0.07	0.14	0.610
Area [Social Science]	0.19	0.16	0.210
Cognitive Engagement	0.06	0.06	0.240
Experiential Attitudes	−0.07	0.06	0.239
Instrumental Attitudes	0.10	0.06	0.124
Extraversion	0.11*	0.05	0.028
Agreeableness	0.03	0.05	0.610
Conscientiousness	0.01	0.05	0.850
Neuroticism	0.09	0.06	0.110
Openness	0.04	0.05	0.467
Random Effects			
Level-two random part:			
τ_{00}			0.893
$\tau_{11\ Task}$			1.427
$\tau_{11\ Method}$			0.056
Level-one variance			
σ^2			0.860
ICC			
N			229
Observations			916

Note: $p < 0.05$, β = Standardized coefficients, β *(SE)*= Standard errors.

4. Discussion

In this study we attempted to investigate the extent to which pre-service teachers agreed on students' grades across different grading approaches (holistic and analytic) and essay quality (i.e., weak vs. strong essay). We also examined the extent to which participants' beliefs about assessment features (receptivity to feedback) and their characteristics (gender, area of specialization, and personality) explained the variability in grade scores. In sum, our analysis demonstrated high intra-individual consistency of teacher grades. That is, teacher candidates provided grades consistently across approaches (holistic and analytic). However, we found that grades showed high variability among participants, independently of the quality of the essay, grading approach, or specialization area, a finding consistent with previous literature (Harsch & Martin, 2013; Jönsson & Balan, 2018; Jönsson et al., 2021). Gender, receptivity to feedback, and personality traits predict the variability in grades among teachers.

4.1. Intra-rater reliability using different grading methods (holistic vs. analytic)

We found that the intra-rater reliability for scores assigned to the four components of the rubric was substantial in the weak essay ($\alpha = 0.762$) and moderate in the strong essay ($\alpha = 0.622$). These findings are consistent with Jonsson and Svingby (2007), who showed that the majority of the studies investigating intra-rater reliability reported alpha coefficients in the range of 0.50 and 0.92, with most values above 0.70. However, a few studies have found very high intra-rater reliability estimates when rubrics were used, whereas others have reported low or moderate estimates (Brookhart, 2018; Johnson et al., 2000; Parkes, 2023). Furthermore, studies have also revealed that, on average, raters overscored low-quality essays and underscored high-quality essays (Eckes, 2008; Leckie & Baird, 2011), which may explain the higher reliability observed for the weak essay in our study. Also, the lack of specific training in the rubric could explain differences in intra-rater reliability, suggesting that judgments about high-quality students' tasks could be more demanding. To this end, Rezaei and Lovorn's (2010) findings showed that teacher candidates tended to be more influenced by mechanical characteristics compared to the content of students' writing, even when they used a rubric. It makes sense as studies have found that few raters receive specialized training in writing instruction and content (Hall, 2016; Hodges et al., 2019) and that differences among raters remained even after rigorous training and calibration methods (Attali, 2016; Eckes, 2008; Engelhard & Myford, 2003).

Our examination of the intra-rater consistency among scores assigned with different grading approaches (holistic vs. analytic) was strong and slightly higher for the weak than for the strong essay, irrespective of participants' area of specialization. Findings indicated that when participants provided grades, they seemed to define their idiosyncratic criteria for assessing students' performance, even when using different methods for assigning grades. Indeed, the correlation between the rubric components and analytic and holistic scores were similar for both essays, with the only difference observed for the organization and

style in the weak essay. In a similar study, Rezaei and Lovorn (2010) noted that raters gave credit to some aspects that were not present in the essay, suggesting that they were influenced by the overall impression regarding some particular aspects of the text instead of trying to provide scores close to the rubric description.

4.2. Inter-rater reliability using different grading methods (holistic vs. analytic)

To answer our second research question, we explored inter-rater reliability across two grading approaches. We found a great variability of grade scores provided by pre-service teachers when assessing student work, regardless of the grading approach and essay quality. The range of scores assigned by teacher candidates using analytic and holistic approaches was virtually identical and equally vast. Some studies garnered similar results. For example, Brimi (2011) identified a range of scores from 50 to 96 on a 100-point scale, even after receiving specific training on how to assess the same written product. Similarly, Rezaei and Lovorn (2010), using a 0 to 100-point scale, found that scores of a wrong essay ranged from 49 to 96 and 32 to 100 for the holistic and analytic grading approaches, whereas for the correct essay, scores ranged from 27 to 98 and 12 and 98, for holistic and analytic approaches, respectively. In other words, the rubric did not appear to help raters to provide more consistent scores.

In our study, we did not train participants on the use of the rubric and observed an alarmingly high range of teacher candidates' scores for both analytic and holistic approaches. A more nuanced examination of our findings showed that using the analytic approach did reduce the range of scores, albeit slightly, compared to the holistic approach. Regarding the inter-rater consistency across the rubric components, results showed a low agreement among pre-service teachers via both Fleiss's Kappa and consensus agreement. Jonsson and Svingby (2007) observed that the Kappa's values reported in the literature varied from 0.20 to 0.63, where values between 0.40 and 0.75 represented fair agreement. In our study, values varied from 0.13 to 0.22.

In terms of the consensus agreement of scores, findings in our sample showed that analytic grades had a higher consensus with expert criteria compared to holistic grades. Interestingly, a higher proportion of participants tended to deviate more from the expert criteria for the strong essay; however, there was more agreement with experts on the weak essay, regardless of the grading approach. Although values are not directly comparable across studies due to variations in the type of participants and tasks, differences in consensus agreement in analytic versus holistic approaches were lower in our study than those reported in other studies. For example, Jönsson et al. (2021) reported a percentage of consensus agreement of 66.7% and 59.7% for analytic and holistic grading conditions, respectively, and Jönsson and Balan (2018) found that analytic condition yielded a substantively higher agreement among assessors as compared to the holistic condition (66% versus 46%).

Our findings were consistent with earlier studies and showed a slight improvement in the consistency when using analytic grades, especially by participants with higher content area expertise (Brookhart & Chen, 2015; Jönsson et al., 2021). The improvement, however, was minimal. Some authors have recommended that a way to enhance rating validity could be through the use of a complementary approach, which combines holistic with analytic scores (Harsch & Martin, 2013; Tomas et al., 2019). Indeed, Tomas et al. (2019) showed evidence that holistic marking practices could be improved by introducing analytic rubrics for feedback as an ancillary during marking.

4.3. Variables that explain variability in grading

To our knowledge, no previous studies have examined individual factors influencing pre-service teachers' grading when using holistic and analytic approaches. Gender, experiential attitudes toward feedback, as well as the personality factor of extraversion, were significant predictors of the likelihood of providing grading scores that agreed with expert criteria and variability in grade scores. However, the magnitude of effects varied across essay quality and grading approaches. Although some experimental studies have found that teachers' gender did not directly relate to differences in grading practices (Doornkamp et al., 2022), we found that male candidates had a higher probability of assigning scores that differed from those of experts using the analytic approach and when rating the weak essay. Moreover, the interaction effect between cohort and gender, suggested that a higher level of domain-specific expertise may help to mitigate existing gender differences in grades. These results are aligned with earlier findings that showed that subject matter expertise provided more accurate differentiation between the qualities of texts or rank orders of student texts (Möller et al., 2022).

Main effects by content expertise, operationalized as area of specialization, were not statistically significant for explaining the probability of agreeing with experts' criteria or the variability in grades. These results are consistent with previous studies that suggested content knowledge does not lead clear differences in teacher judgments (Jansen et al., 2021; Meadows & Billington, 2010). The most compelling explanation for the present set of findings is that novel teachers' ability to provide accurate judgments about students' writing essays is not solely dependent on their accumulated content knowledge in the subject being assessed. Teacher candidates, regardless of their area of specialization, are in the process of developing and internalizing the standards, methods, and formalisms of the teaching profession. Consequently, their judgments may be influenced more by idiosyncratic aspects of the evaluation process and a simplified, common view of the writing task assessment. This suggests that the variability in grading may stem from the pre-service teachers' evolving understanding and application of assessment criteria rather than a lack of expertise in the content area (Jansen et al., 2020).

Regarding personality, our results showed that extraversion was positively associated with assigning higher grades. This finding aligns with the idea that extraverted individuals, who are generally more sociable, energetic, and optimistic, may tend to provide more favorable evaluations. However, previous empirical evidence on the impact of personality traits on teachers' judgments and grading practices is extremely limited. This gap in the literature underscores the need for further investigation. Understanding the influence of personality traits on grading can provide valuable insights into how subjective factors may affect educational outcomes. Therefore, we strongly encourage researchers to explore this intriguing avenue of research to better comprehend the nuances of grading practices and to develop strategies that ensure fairness and objectivity in educational assessments.

4.4. Limitations and future directions

Although this exploratory study provides meaningful information about the importance of teacher's variables to understand differences in grading variability and rationality, this study is not without limitations. First, we only included a sample of pre-service teachers in a single institution. Hence, our results should be cautiously generalized to other populations of teacher candidates. Future studies could explore individual and contextual factors that influence teacher candidates' and experienced teachers' grading decisions in diverse populations at different educational stages. Second, the sterile study conditions differed from typical classroom assessments. More specifically, the participants did not design the task themselves and had no personal relationship with students. Therefore, researchers could replicate our study in situ. Third, in our sample, teacher candidates did not receive specific training on the rubric to provide analytic scores. Although multiple studies about human essay rating suggested that even though rigorous training and calibration methods are employed, raters still differ in their scores (e.g., Attali, 2016), the absence of training could be associated with overestimation in the variability of the analytic condition. Therefore, future

studies may examine the extent to which training candidates affect variability in grading decisions.

Finally, one of the main limitations of this study is the potential impact of the sequence of written essay and the anchor effects of grading outcomes. In the first cohort (2020), neither the order of essay quality (strong essay followed by weak essay) nor the grading approach (holistic followed by analytic) was counterbalanced. This lack of counterbalancing could have influenced the results, as it is well-documented that the quality of the first essay can significantly affect the rating of the subsequent essay (Steiner & Rain, 1989). Similarly, sequence effects are known to occur with different grading methods, where holistic approach can impact subsequent analytic ratings (Klein et al., 1998). These effects could potentially inflate the inter-rater reliabilities, affecting the study's outcomes. In the second cohort (2021), we addressed this by randomizing the order of essay quality presented to participants, although the sequence of grading methods remained the same as in the first cohort. Our results thus showed slightly better indicators for weak essays across both samples, suggesting that the lack of counterbalancing of essays does not skew our findings. However, the influence of the initial grading score (holistic) as a confounder of subsequent evaluation (analytic) cannot be entirely ruled out, although our results regarding intra-rater reliability were consistent with previous literature.

Future research should prioritize fully counterbalancing both the order of essay presentation and the sequence of grading methods to mitigate these potential confounding variables. Additionally, exploring alternative designs that minimize sequence effects would provide a more robust understanding of the differences between holistic and analytic grading approaches. Another promising direction for future studies is to investigate how different grading sequences might influence the development of pre-service teachers' evaluative skills over time, offering insights into optimizing grading practices in educational settings.

4.5. Practical implications and conclusions

In this study, we examined pre-service teachers' grades to writing task assigned using analytic and holistic approaches and explored how their beliefs about assessment features and personal characteristics explained variability in grading practices. Although findings are specific to the assessment of writing and cannot be generalized to another disciplines, such as mathematics or science, our research revealed intriguing patterns that warrant further investigations. Our findings showed high discrepancy among participants-assigned grades for both analytic and holistic approaches. Conversely, we found a high intra-individual consistency when scoring essays using both grading methods. Hence, our results suggest that within the same community of practices, pre-service teachers weigh the criteria for providing grades differently, thus generating great variability when assessing student work. This pattern was observed even though the criteria to provide scores used a limited number of parameters and teachers did not have access to students' non-cognitive factors that could bias their judgments. Results also provide insight regarding individual characteristics that should be considered when exploring variability in grading scores.

In terms of its practical value, this study's most important implication is the need to interpret grades with caution. Different teachers from the same school would have a high chance of assigning different scores to the same assignment. Hence, constant discussions within schools or departments should take place. Further, some variables that explain the variability and lack of consensus with experts, such as receptivity of instructional feedback, represent a relatively malleable characteristic that could be influenced through interventions during teacher training programs (Winstone et al., 2017). Moreover, receiving high-quality feedback could impact raters' behavior and assist them in assigning accurate scores (Wendler et al., 2019). Therefore, future training courses designed for teacher candidates could emphasize providing adequate and effective feedback that enhances grading practices. Considering teachers' gender, discipline, and beliefs about features of assessment is also important. In sum, teacher candidates deserve more training to reduce the variability of grades, and whether it is done through holistic or analytic grading, it does not appear to matter at all.

Funding

This work was not funding supported for any organization.

Open Science statement

Open Science: We report all data exclusions, all data exclusion criteria (if any), whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If I use inferential tests, I report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including the codebook and supplementary materials can be retrieved from https://osf.io/2j6ah/https://osf.io/2j6ah/?view_only=bc98b7abb8e7459ba55c85311f330cad.

Open Materials: The information needed to reproduce all of the reported methodology is not openly accessible. The material is available on request from author(s). Preregistration of Studies and Analysis Plans: This study was not preregistered.

CRediT authorship contribution statement

Carolina Lopera-Oquendo: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Anastasiya A. Lipnevich:** Conceptualization, Supervision, Writing – review & editing. **Ignacio Mañez:** Data curation, Writing – review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2024.101992.

References

Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record, 103*, 760–822. https://doi.org/10.1111/0161-4681.00134

Attali, Y. (2016). A comparison of newly trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54–74.

Barrick, M. R., & Mount, M. K. (1991). The Big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Bastian, K. C., McCord, D. M., Marks, J. T., & Carpenter, D. (2015). *Do Personality Traits Impact Beginning Teacher Performance and Persistence?*. University of North Carolina. https://www.wcu.edu/webfiles/pdfs/ceap_personalitytraits_2015.pdf.

Bastian, K. C., McCord, D. M., Marks, J. T., & Carpenter, D. (2017). A temperament for teaching? Associations between personality traits and beginning teacher performance and retention. *AERA Open, 3*(1). https://doi.org/10.1177/2332858416684764

Bean, G. J., & Bowen, N. K. (2021). Item Response Theory and Confirmatory Factor Analysis: Complementary Approaches for Scale Development. *Journal of Evidence-Based Social Work, 18*(6), 597–618. https://doi.org/10.1080/26408066.2021.1906813.

Benet, V., & John, O. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729–750. https://doi.org/10.1037//0022-3514.75.3.729

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate Grade Point Average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources, 34*(2), 268–293.

Blömeke, S., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G., & König, J. (2015). Teacher change during induction: Development of beginning primary teachers' knowledge,

beliefs and performance. *International Journal of Science and Mathematics Education, 13*, 287–308. https://doi.org/10.1007/s10763-015-9619-4
Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655–670. https://doi.org/10.1080/03075071003777716
Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607
Bonner, S. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 87–106). SAGE Publications. https://doi.org/10.4135/9781452218649.n6.
Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences, 113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113
Bouwer, R., Verhavert, S., Lesterhuis, M., Van Gasse, R., Donche, V., & De Maeyer, S. (2017). Interpreting the validity of misfit statistics in Comparative Judgement. In *AEA-europe conference on assessment Cultures in a globalized world, Prague, Czech republic*.
Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educ. Res. Eval., 17*, 141–159. https://doi.org/10.1080/13803611.2011.597112
Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of Educational Research, 105*, 176–195. https://doi.org/10.1080/00220671.2011.552075
Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *High School Journal, 96*, 77–100. https://doi.org/10.1353/hsj.2013.0000
Brimi, H. M. (2011). Reliability of grading high school work in English. *Practical Assessment, Research and Evaluation, 16*(17), 1–12. https://doi.org/10.7275/j531-fz38
Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170
Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*. https://www.frontiersin.org/articles/10.3389/feduc.2018.00022.
Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565
Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. https://doi.org/10.7916/D8NV9JQ0
Brookhart, S. M., & Nitko, A. J. (2014). *Educational assessment of students*. Limited: Pearson Education.
Camara, W. J., & Echternacht, G. (2000). *The SAT[R] I and high school grades: Utility in predicting success in College* (pp. 10023–16992). New York: Research Notes. The College Board. https://eric.ed.gov/?id=ED446592.
Cheng, L., & Sun, Y. (2015). Teachers' grading decision making: Multiple influencing factors and methods. *Language Assessment Quarterly, 12*(2), 213–233. https://doi.org/10.1080/15434303.2015.1010726
Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*(2), 159. https://doi.org/10.1207/s15326977ea0302_3
Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources, 48*(1), 236–264. https://www.jstor.org/stable/23799113.
Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53–72. https://doi.org/10.1207/s15324818ame1201_4
Doornkamp, L., Van der Pol, L. D., Groeneveld, S., Mesman, J., Endendijk, J. J., & Groeneveld, M. G. (2022). Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs. *Teaching and Teacher Education, 118*, Article 103826. https://doi.org/10.1016/j.tate.2022.103826
Duncan, R. C., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research, 53*, 1–21.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155–185. https://doi.org/10.1177/0265532207086780
Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and Composition program with a Many-Faceted Rasch Model. *Research report No. 2003-1. ETS RR-03-01. College Board, NY*. https://eric.ed.gov/?id=ED561016.
Federičová, M. (2015). Gender gap in application to selective schools: Are grades a good signal? *CERGE-EI Working Paper Series No. 550*. https://doi.org/10.2139/ssrn.2685192
Fulmer, G., Lee, I., & Tan, K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assessment in Education: Principles, Policy & Practice, 22*, 1–20. https://doi.org/10.1080/0969594X.2015.1017445
Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34. https://doi.org/10.1037/0003-066X.48.1.26
Guskey, T. R., & Bailey, J. M. (2010). *Developing standards-based report cards*. Thousand Oaks, CA: Corwin Press.
Guskey, T. R., & Link, L. J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303–320. https://doi.org/10.1080/0969594X.2018.1555515
Hall, A. H. (2016). Examining shifts in preservice teachers' beliefs and attitudes toward writing instruction. *Journal of Early Childhood Teacher Education, 37*(2), 142–156. https://doi.org/10.1080/10901027.2016.1165761
Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*(4), 146–168.
Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422
Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Sturmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*, 1e24. https://doi.org/10.14786/flr.v7i4.384
Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*, 682–690.
Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K., & McTigue, E. (2019). Developing and examining validity evidence for the writing rubric to inform teacher educators (WRITE). *Assessing Writing, 40*, 1–13. https://doi.org/10.1016/j.asw.2019.03.001
Isnawati, I., & Saukah, A. (2017). Teachers' grading decision making. *Teflin Journal - A publication on the teaching and learning of English, 28*, 155. https://doi.org/10.15639/teflinjournal.v28i2/155-169
Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/0146621616664046
Jansen, T., Machts, N., Vögelin, C., Keller, S., & Möller, J. (2020). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*. https://doi.org/10.1016/j.tate.2020.103216
Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*, Article 103216. https://doi.org/10.1016/j.tate.2020.103216
Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*, 121–138. https://doi.org/10.1207/S15324818AME1302_1
Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974
Jönsson, A., & Balan, A. (2018). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research and Evaluation, 23*(12), 1–11. https://doi.org/10.7275/mg59-xq60
Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice, 28*(3), 212–227. https://doi.org/10.1080/0969594X.2021.1884041
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002
Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875–925. https://doi.org/10.1037/a0033901
Kim, L. E., Dar-Nimrod, I., & MacCann, C. (2018). Teacher personality and teacher effectiveness in secondary school: Personality predicts teacher support and student self-efficacy but not academic achievement. *Journal of Educational Psychology, 110*(3), 309–323. https://doi.org/10.1037/edu0000217
Kim, L. E., Jörg, V., & Klassen, R. M. (2019). A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational Psychology Review, 31*(1), 163–195. https://doi.org/10.1007/s10648-018-9458-2
Klapp, A. (2016). The importance of self-regulation and negative emotions for predicting educational outcomes – evidence from 13-year olds in Swedish compulsory and upper secondary school. *Learning and Individual Differences, 52*, 29–38. https://doi.org/10.1016/j.lindif.2016.10.013
Klassen, R. M., Durksen, T., Kim, L., Patterson, F., Rowett, E., Warwick, J., … Wolpert, M.-A. (2017). Developing a Proof-of-Concept Selection Test for Entry into Primary Teacher Education Programs. *International Journal of Assessment Tools in Education, 4*(2), 96–114. http://ijate.net/index.php/ijate/article/view/136.
Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76. https://doi.org/10.1016/j.edurev.2014.06.001
Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., … Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121–137. https://doi.org/10.1207/s15324818ame1102_1
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7
Kunnath, J. P. (2017). Teacher grading decisions: Influences, rationale, and practices. *American Secondary Education, 45*, 68–88.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009

Lavy, V., & Sand, E. (2016). On the origins of gender human capital gaps: Short- and long-term consequences of teachers' stereotypical biases. *IDEAS Working Paper Series from RePEc*. https://www.proquest.com/publiccontent/docview/2185158758?pq-origsite=primo.

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation, 14*(2), 181–199. https://doi.org/10.1080/13803610801956663

Lekholm, A. K., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*(1), 1–23. https://doi.org/10.1080/13803610802470425

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

Lindahl, E. (2016). Are teacher assessments biased? – Evidence from Sweden. *Education Economics, 24*(2), 224–238. https://doi.org/10.1080/09645292.2015.1014882

Loibl, K., Leuders, T., & Doĕrfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91*, Article 103059. https://doi.org/10.1016/j.tate.2020.103059

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in Mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78–102. https://doi.org/10.1080/10627190903039429

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*, 20–32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203–213. https://doi.org/10.1080/00220670209596593

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. *Metropolitan Educational Research Consortium, Richmond, VA*. https://eric.ed.gov/?id=ED447195.

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: AQA Centre for Education Research and Policy.

Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., & Reble, R. (2022). Judgment accuracy of German student texts: Do teacher experience and content knowledge matter? *Teaching and Teacher Education, 119*, Article 103879. https://doi.org/10.1016/j.tate.2022.103879

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_9.

Parkes, J. (2023). Reliability in classroom assessment. In J. H. MacMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 107–123). SAGE Publications. https://doi.org/10.4135/9781452218649.

Pliske, R., & Klein, G. (2003). The naturalistic decision-making perspective. In S. L. Schneider, & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 559–585). Cambridge University Press. https://doi.org/10.1017/CBO9780511609978.019.

Protivínský, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies In Educational Evaluation, 59*, 141–149. https://doi.org/10.1016/j.stueduc.2018.07.006

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis, 42*(3), 375–392. https://doi.org/10.3102/0162373720932188

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research, 102*(3), 175–185. https://doi.org/10.3200/JOER.102.3.175-186

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*(7), 1372–1380. https://doi.org/10.1016/j.tate.2010.03.008

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Read, B., Francis, B., & Robson, J. (2005). Gender, "bias", assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education, 30*(3), 241–260. https://doi.org/10.1080/02602930500063827

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82*(1), 30–43. https://doi.org/10.1037/0021-9010.82.1.30

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*(3), 323–346. https://doi.org/10.1348/096317903769647201

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 1–97. https://doi.org/10.1007/BF03372160

Sanrey, C., Bressoux, P., Lima, L., & Pansu, P. (2021). A new method for studying the halo effect in teachers' judgement and its antecedents: Bringing out the role of certainty. *British Journal of Educational Psychology, 91*(2), Article e12385. https://doi.org/10.1111/bjep.12385

Simonton, D. K. (2003). Expertise, competence, and creative ability: The perplexing complexities. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (p. 213e238). Cambridge University Press. https://doi.org/10.1017/CBO9780511615801.010.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136–142. https://doi.org/10.1037/0021-9010.74.1.136

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9*, 1–19. https://doi.org/10.7275/96jp-xz07

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00089

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education*. https://doi.org/10.3389/feduc.2019.00089/full

Tomlinson, C. A. (2000). Differentiation of instruction in the elementary grades. *ERIC Digest. ERIC Clearinghouse on Elementary and Early Childhood Education, Champaign, IL*. https://eric.ed.gov/?id=ED443572.

Wendler, C., Leusner, D., Thompson, V., & Tolentino, F. (2019). *Examining the perceived effectiveness of rater feedback (Research Memorandum No. RM-19-13)*. Princeton, NJ: Educational Testing Service.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*(1), 21–39. https://doi.org/10.1007/s10212-014-0225-6

Carolina Lopera-Oquendo is a doctoral student and graduate fellowship in Educational Psychology cat Graduate Center, City University of New York, NY. Her research interests are instructional feedback, measurement, evaluation and assessment in educational settings.

Dr. Anastasiya Lipnevich is gy at Queens College and The Graduate Center, City University of New York, NY.

Dr. Ignacio Manez is an assistant professor of Department of Evolutive Psychology and Education at Universitat de Valencia, Spain