Humanities and Social Sciences Communications

Article in Press

https://doi.org/10.1057/s41599-026-07862-0

The mediating role of validity assessment in EFL teachers' cognitive development within writing assessment validity arguments

Received: 26 November 2024

Accepted: 28 May 2026

Cite this article as: Ke, Y., Chen, L., Wang, G. *et al.* The mediating role of validity assessment in EFL teachers' cognitive development within writing assessment validity arguments. *Humanit Soc Sci Commun* (2026). https://doi.org/10.1057/s41599-026-07862-0

Yuguo Ke, Liang Chen, Gang Wang & Xiaozhen Zhou

The mediating role of validity assessment in EFL teachers' cognitive development within writing assessment validity arguments

Abstract

Although the need to establish validity in writing assessment is widely acknowledged—particularly given persistent challenges such as scoring subjectivity, construct underrepresentation, ambiguous criteria, and inadequate rater training—research remains notably limited regarding how EFL teachers' cognitive development mediates their assessment competence. This cross-sectional study introduces an innovative mediation approach grounded in the validity argument framework to explore how writing assessments are validated in practice. The findings indicate that teacher cognition—specifically the cognitive processes underlying evaluative reasoning and rubric interpretation—mediates the relationship between EFL teachers' assessment literacy and their scoring practices. These results highlight the potential for designing targeted interventions to foster cognitive growth among EFL teachers. By integrating computational modeling into validity argumentation, future research and pedagogical practices can more effectively enhance teachers' assessment literacy, ultimately contributing to more valid, consistent, and instructionally meaningful writing assessments in EFL contexts.

Introduction

As educational assessment continues to evolve, the validity of EFL writing evaluation has attracted growing scholarly attention (Cronbach, 1988; Kunnan & Jang, 2009; Sireci, 2013; Correnti et al., 2022; Sawaki, 2024; Romig et al., 2025). The inherently complex and multidimensional nature of writing assessment in EFL contexts — often characterized by overgeneralization, vagueness, subjectivity, and inconsistency — raises persistent concerns regarding fairness, reliability, and scoring accuracy (Bachman & Palmer, 1996; Taylor et al., 2023; Thwaites et al., 2025). These concerns are particularly salient in exam-oriented educational systems, where writing assessments carry high stakes, underscoring the need for EFL teachers to deeply understand and effectively apply validity principles (Cronbach, 1988; Ferrara & Qunbar, 2022).

Within this context, EFL teachers' cognitive development has emerged as a key mediator of assessment validity (North & Piccardo, 2016). Although foundational knowledge remains essential, contemporary research emphasizes that understanding the cognitive processes underlying assessment decisions is equally critical. Accordingly, this study examines both teachers' assessment literacy (their knowledge base) and their teacher cognition (the cognitive processes they employ) to develop a comprehensive understanding of assessment behavior. Together, these dimensions shape teachers' capacity to make sound evaluative judgments, ensure scoring consistency, and promote fairness. Teacher cognition equips educators with the interpretive tools necessary to critically engage with validity arguments in authentic classroom environments (Cronbach, 1988; Folger et al., 2023).

From the perspective of validity argumentation, enhancing the quality of writing assessments has become a central goal of broader assessment reform. Contemporary research increasingly identifies atypical rater behavior—rather than issues inherent solely to task or prompt design—as a significant

threat to validity. Although these insights have spurred progress in reducing subjectivity and refining scoring procedures, many studies have yet to address deeper structural challenges, particularly those related to the operationalization, implementation, and pedagogical integration of validity argument frameworks.

Despite growing recognition of the relationship between assessment literacy and the quality of writing assessment, few empirical studies have proposed or validated comprehensive models that illustrate how validity arguments are constructed and enacted in practice. Limited attention has been devoted to how abstract validity principles translate into concrete assessment design, scoring practices, and instructional decision-making. Consequently, our understanding of how these frameworks function in real-world classroom settings—especially within East Asian educational systems such as those in China, Japan, and Korea—remains underdeveloped.

In this study, we reframe the research problem to clarify that our focus is on how teachers develop validity argument cognition for use in daily classroom practice, particularly in formative assessment. We explicitly distinguish this focus from validity argumentation for large-scale standardized testing, drawing on relevant literature (e.g., Bachman & Palmer, 2010; Chapelle, 2021). Given the complex interrelationships among teacher cognition, assessment practices, and scoring outcomes, there is a pressing need for research that investigates the mediating mechanisms linking these domains. Findings from such inquiry could inform the design of targeted, evidence-based professional development programs aimed at enhancing EFL teachers' teacher cognition—defined by Borg (2003) as "the complex, practically oriented, personalized, and context-sensitive networks of knowledge, thoughts, and beliefs that language teachers draw on in their work."

Ultimately, the fairness and reliability of writing assessments in early educational stages are closely tied to teachers' ongoing professional development. Providing educators with context-sensitive interventions that strengthen their interpretive and evaluative capacities is essential. By building on existing strengths and addressing persistent gaps, we can foster more valid, equitable, and pedagogically meaningful writing assessment practices across diverse EFL settings.

Theoretical framework

A growing body of research highlights the strong relationship between writing assessment literacy and the validity of writing evaluations. Assessment literacy refers to the knowledge base—the "what" that teachers know—including theoretical knowledge of assessment, familiarity with scoring procedures, and understanding of assessment purposes, as captured by our adapted questionnaire. It represents the declarative and procedural knowledge a teacher possesses (Kim, 2020). Teacher cognition, in contrast, refers to the cognitive processes—the "how" of teachers' thinking, reasoning, and decision-making (Borg, 2003). This encompasses the dynamic, in-the-moment mental activities teachers engage in when applying their assessment knowledge, including evaluative reasoning, metacognitive analysis, and the cognitive mediation strategies outlined in our theoretical framework.

Studies consistently show that raters who receive formal training in assessment are more likely to apply validity principles consistently in their mediating behaviors (Christie, 2002). By contrast, educators without such training often rely on less structured and empirically unsupported scoring methods (Council of Europe, 2020; Ferrara & Qunbar, 2022). Moreover, the ability to formulate and apply validity arguments has been linked to both cognitive development and teaching

experience. Insights from cognitive science further support this connection, underscoring that evidence-based reasoning plays a critical role in the interpretive processes involved in evaluation and mediation (Piantadosi et al., 2016; Sanchis & Viruega, 2024; Choi et al., 2025). Consequently, increasing scholarly attention has turned to the cognitive beliefs and mediating mechanisms that influence how teachers interpret rubrics, deliver feedback, and form scoring judgments (Fulcher, 2010; McNamara & Knoch, 2019; Dorsey & Michaels, 2022; Thwaites & Paquot, 2025). The present study considers two functions related to relational mediation (Christie, 2002) and four functions associated with cognitive mediation (Nadal & Thome, 2021).

Within developmental psychology, greater emphasis is now placed on specific teacher attributes—particularly familiarity with validity argumentation—and their impact on assessment practices. This focus on cognitive mediation has prompted investigations into the alignment between teachers' internal reasoning processes and their observable scoring behaviors. For example, Sawaki and Koizumi (2017) examined the extent to which raters' cognitive mediation is reflected in their actual scoring decisions. Other studies have explored related areas, such as standard-setting procedures, inter-rater reliability, and calibration strategies (Dechter et al., 2013).

Despite these advances, several nuanced aspects of teacher cognition remain underexplored. These include how educators interpret and apply scoring rubrics (Chapelle et al., 2008; Wollenschläger et al., 2016), how they establish instructional and assessment goals (Knoch, 2009; Knoch & Chapelle, 2018; Stephens & Sarkar, 2025), and how they conceptualize fairness, generalizability, and accountability in high-stakes writing contexts (Bachman & Palmer, 2010; Lallmamode et al., 2016; Thwaites et al., 2025). Equally important are issues related to feedback—such as its clarity, coherence, and alignment with assessment purposes—which are crucial for effective formative assessment and student learning progress (Arter & McTighe, 2001; Panadero & Jonsson, 2013; Ghaffar et al., 2020; Sawaki, 2024).

Recent research employing the Rubric Use Argument (RUA) framework has provided new insights into how teachers conceptualize and implement notions of quality in writing assessment (Toulmin, 1958, 2003; Jonsson & Svingby, 2007; Fulcher, 2010; Castillo et al., 2023; Saeli & Rahmati, 2023). These studies indicate that teachers with more advanced cognitive development demonstrate more consistent and nuanced rubric application, thereby enhancing assessment accuracy and validity. Importantly, research on rubric-based assessment also cautions against attributing scoring inconsistencies solely to teacher-related factors. Instead, variability in evaluation outcomes is often tied to ambiguities or flaws in rubric design, which can obscure scoring criteria and compromise reliability (Bachman & Palmer, 2010; Fulcher, 2010; Correnti et al., 2022; Castillo et al., 2023; Sawaki, 2024; Ke, 2024).

Further evidence points to a reciprocal relationship: teachers working within high-validity assessment environments tend to adapt their cognitive strategies to align with evidence-based evaluative practices (Fulcher, 2010). A recent systematic review by Wan et al. (2019) outlines distinguishing traits between high- and low-validity teacher groups, emphasizing the role of teaching experience, assessment orientations, and feedback methodologies in shaping teachers' approaches to student writing. In line with the focus of this study, we have tailored the validity argument model for classroom-based, formative assessment purposes. Accordingly, we have adjusted the language throughout this section to emphasize claims relevant to teachers' instructional decisions—such as the meaningfulness of assessment tasks for learning and the consistency of

scoring in classroom contexts—rather than claims more typical of large-scale test validation, such as population generalizability.

Together, these findings underscore the essential mediating role of teacher cognition in determining the quality, consistency, and pedagogical usefulness of writing assessments in EFL contexts (see Fig. 1). In this model, teacher cognition functions as the mediating mechanism through which a teacher's assessment knowledge is translated into action. A teacher may possess high assessment literacy—a rich knowledge base—but it is their cognitive processes—their ability to interpret a rubric, reason through a student's response, and make a mediated judgment—that determine how that knowledge is actually applied in the classroom. Therefore, teacher cognition logically mediates the relationship between assessment literacy and assessment outcomes. Teacher cognition is not a component of assessment literacy; rather, it is the process that operationalizes it.

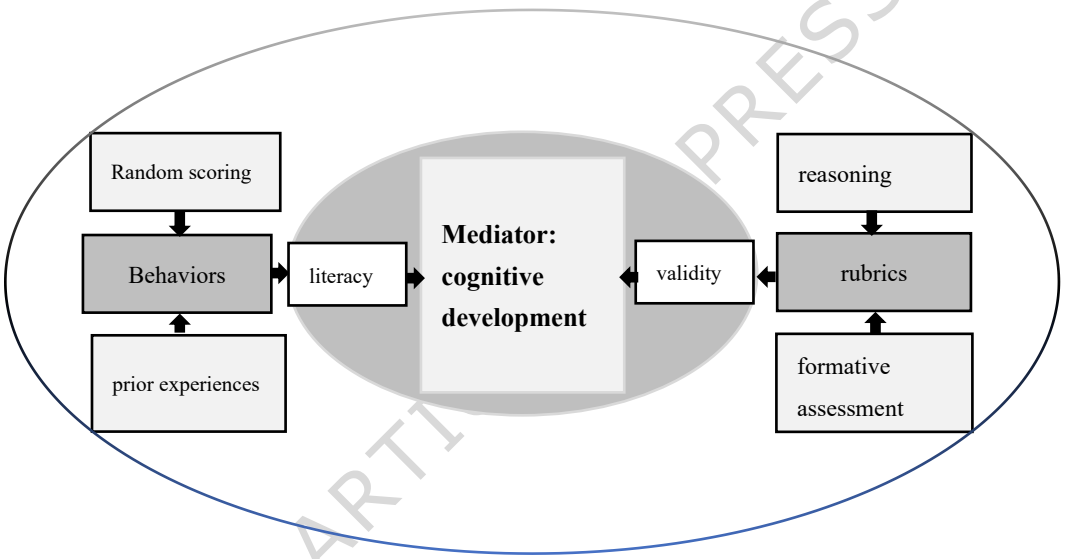

Fig.1 Hypothesized relationships between the assessment literacy and validity argument

The present study

This study aimed to investigate the mediating role of EFL teachers' cognitive development in their construction of validity arguments within writing assessment. Specifically, we examined whether the cognitive dimensions of assessment literacy mediate the relationship between teachers' assessment-related characteristics and the overall quality of their writing evaluations. The research focused on a targeted group of EFL teachers who exhibited early signs of validity-related inaccuracies—an understudied yet critical issue with direct consequences for the fairness, reliability, and credibility of writing assessment outcomes.

Although previous research has established broad correlations between assessment literacy and evaluation quality, few studies have explored how teachers cognitively engage with validity arguments themselves (Correnti et al., 2022). Notably, the cognitive and pedagogical processes underlying this engagement may differ substantially from those observed in the general teaching population (Toulmin, 1958, 2003). Participants in this study were recruited from an exam-oriented professional development program and may therefore face more pronounced challenges in assessment reasoning compared to teachers with formal training in language assessment.

Building on existing literature, we hypothesized that heightened assessment-related traits—particularly those manifested through complex cognitive behaviors during scoring—may disrupt the mediating influence of cognitive development on the principled application of validity arguments. Such interference could undermine the accuracy and consistency of teachers' validity judgments, especially given the central role of writing tasks in high-stakes language assessments. The research questions guiding this study are as follows:

Research Question 1: What mediating roles does EFL teachers' cognitive development play in the construction of validity arguments in writing assessment?

Research Question 2: To what extent do indicators of teaching experience mediate teachers' writing rating behavior within validity argumentation?

We further hypothesized that limited comprehension or misinterpretation of fundamental assessment constructs would lead to a misalignment between scoring practices and underlying validity principles. Such cognitive dissonance could manifest as inconsistent or unreliable evaluation behaviors. Correspondingly, we proposed that a stronger mediating effect of cognitive development, along with more informed assessment attitudes, would be positively associated with teachers' engagement in validity-based reasoning. Importantly, we posited that this relationship is mediated by teachers' cognitive interaction with validity argument frameworks—a mechanism that may help explain observed patterns of inaccuracy in EFL writing assessment.

Materials and Methods

Study design and participants. A total of 210 EFL teachers participated in this study. They were categorized into seven groups based on their years of teaching experience, a stratification that ensured diverse representation across varying levels of professional expertise and instructional contexts. Participants were recruited through two major national academic events in language education: the National Language Testing Conference and the National Applied Linguistics Conference. Following initial referral, all candidates completed the Teacher Writing Assessment Literacy (TWAL) checklist—a 12-item instrument conceptually developed by operationalizing Toulmin's (1958, 2003) model of argumentation. Toulmin's framework was adopted because teacher scoring can be conceptualized as an internal argumentation process, and its components provide a structured lens for analyzing the cognitive processes underlying scoring decisions within validity arguments.

Due to the heterogeneity of recruitment sources, a member of the research team conducted individual eligibility screenings to ensure consistent participant inclusion. To qualify, individuals were required to exhibit "atypical" responses on at least three of five core TWAL items. These items, empirically validated by Chan et al. (2023), have been identified as strong predictors of cognitive challenges in constructing validity arguments in writing assessment. The five key indicators captured behavioral tendencies such as: reliance on subjective or intuition-based rating approaches; inconsistent or overgeneralized scoring practices; vague or non-standardized rubric application; and limited awareness or neglect of foundational validity principles. The selection of these indicators was informed by prior literature on writing assessment validation (e.g., Kane, 2013) and by preliminary findings from our pilot study.

The TWAL checklist was designed to identify behavioral markers of cognitive mediation during assessment-related tasks. Each item required teachers to perform a specific writing assessment task,

with trained raters evaluating their responses as either "typical" or "atypical" based on alignment with validity-informed reasoning. For example, one item assessing teachers' adherence to standardized criteria stated: "Rate the writing samples using a five-point scale based on standard writing criteria. Consider the structural components of rating and refer to AMOS-based analyses of writing assessment structures." According to Fulcher (2010), the TWAL instrument demonstrates a predictive validity rate of 72% in identifying educators likely to experience cognitive challenges related to writing assessment accuracy. This establishes the TWAL as a reliable diagnostic tool for detecting early indicators of validity-related inaccuracies in rater behavior, particularly within EFL teaching contexts, thereby supporting the investigation of cognitive mediation in validity argument construction.

Instruments

The data for this study were collected using two primary instruments: the Chinese English Teaching Writing Assessment Competence Corpus (CETWACC) and an adapted version of the Teacher Writing Assessment Literacy (TWAL) questionnaire. Together, these instruments were designed to capture both the demonstrated assessment competence and self-reported assessment literacy of EFL teachers, with a particular focus on the cognitive dimensions underlying validity argumentation in writing evaluation.

The framework adopted in this study serves as a comprehensive evaluative tool for assessing EFL teachers' competence in writing assessment, focusing specifically on two core dimensions: teaching experience and instructional level. These dimensions were selected based on substantial empirical evidence underscoring the influential roles that both length of teaching practice and educational stage (e.g., primary, secondary, tertiary) play in shaping teachers' cognitive engagement with assessment tasks, evaluative behaviors, and comprehension of validity principles (Bachman, 2005; Fulcher, 2010; Wu, 2025). To enable systematic and nuanced appraisal, the framework—designated as the Cognitive Evaluation of Teachers' Writing Assessment Competence and Calibration (CETWACC)—categorizes teacher performance across 60 distinct sub-dimensions. These sub-dimensions emerge from the intersection of seven bands of teaching experience and six instructional level categories (spanning from elementary to higher education). This detailed classification facilitates a granular examination of how cognitive competence in writing assessment evolves across diverse career stages and instructional settings.

Each sub-dimension is evaluated using a four-point ordinal scale ranging from 0 (no demonstrated competence) to 3 (highly competent), guided by explicit performance descriptors. These descriptors assess both the depth of cognitive mediation during assessment tasks and the degree of alignment with validity-oriented principles. Scoring was conducted by expert raters specializing in language testing and teacher education. The scoring procedure attained strong inter-rater reliability (Cohen's $\kappa = 0.82$), ensuring consistency and objectivity across varied teacher profiles.

Scores from all 60 sub-dimensions were aggregated to compute a composite CETWACC score, with a maximum possible total of 100. This composite serves as an integrated measure of a teacher's assessment competence, reflecting both instructional context and professional development trajectory. For instance, an experienced secondary-level teacher may exhibit strong competence in areas such as rubric application, feedback coherence, and scoring consistency,

whereas a novice primary teacher might demonstrate developing skills—particularly in applying validity concepts and interpretive judgment.

Notably, the CETWACC framework does not presume a linear progression in assessment competence. Instead, it accommodates non-linear developmental pathways, acknowledging that extensive teaching experience does not invariably lead to advanced assessment expertise. Similarly, early-career teachers may display notably high competence due to targeted training, reflective practice, or active participation in professional learning communities. By accounting for such variability, the framework captures both typical and atypical cognitive profiles—a crucial capacity for supporting the validity argument analyses central to this study.

Teacher Writing Assessment Literacy (TWAL) Questionnaire. In this study, we adapted and modified the original Language Assessment Literacy (LAL) scale developed by Ke (2023) to better align with the specific objectives and contextual focus of our research on writing assessment competence. The original instrument comprises 53 items organized into ten core dimensions: Theoretical Knowledge (5 items), Assessment Principles and Concepts (4 items), Language Pedagogy (9 items), Impact and Social Values (4 items), Local Assessment Practices (6 items), Personal Beliefs and Attitudes (4 items), Scoring and Decision-Making (5 items), Assessment Construction (6 items), Administration and Scoring Procedures (5 items), and Assessment Evaluation (5 items).

To enhance the scale's relevance to the four key challenges in writing assessment—namely, overgeneralization, vagueness, subjectivity, and randomness, as originally identified by Cronbach (1988) and further elaborated by Ferrara and Qunbar (2022)—we conducted a comprehensive review and introduced targeted revisions. Building on this established framework, we analyzed participants' questionnaire responses and writing assessment outcomes using a categorical approach structured around these four dimensions. Revisions included rewording existing items for greater clarity and contextual relevance, as well as developing new items that reflect the pedagogical aims, procedural practices, and contextual realities of writing assessment within our research setting—with particular attention to mitigating the four aforementioned weaknesses.

Special emphasis was placed on aligning the scale with local curriculum objectives, region-specific assessment practices, and contemporary theoretical advances in writing assessment validity. To strengthen both construct and content validity, we incorporated additional items designed to directly probe teachers' understanding of validity theory, core validity principles, and their practical application in writing assessment contexts. The revised instrument also integrates cognitive aspects of assessment literacy, including teacher beliefs, evaluative dispositions, and decision-making strategies that underlie the formation of validity arguments.

All new and modified items underwent a rigorous expert review process to ensure conceptual clarity, contextual appropriateness, and consistency with the study's overarching goals. To guarantee linguistic and cultural accuracy, the scale was translated into the target language and back-translated by bilingual experts specializing in language assessment and teacher education. The lead coder and a second researcher independently coded a subset of 20% of the responses, followed by a discussion session to refine code definitions and resolve initial disagreements through consensus; when consensus could not be reached, a third researcher was consulted.

The finalized version of the adapted LAL scale exhibited high internal consistency, with a

Cronbach's alpha of 0.92 for the full instrument, indicating strong reliability. Participants rated each item on a five-point Likert scale ranging from 1 (Not at all Knowledgeable/Skilled) to 5 (Extremely Knowledgeable/Skilled), allowing for a nuanced evaluation of their self-perceived competence in writing assessment literacy.

An experienced secondary-level teacher might receive a score of 3 (highly competent) on sub-dimensions related to rubric application, demonstrating the ability to interpret ambiguous criteria by linking them to specific textual features and instructional goals. In contrast, a novice primary teacher might receive a score of 1 (developing) on sub-dimensions related to validity concept application, reflecting emerging but inconsistent use of validity principles in evaluative reasoning. One revised item in the Assessment Principles and Concepts dimension reads: "I can identify instances of overgeneralization in my scoring criteria and adjust my judgments to ensure they accurately reflect the specific construct being assessed." Participants rate their perceived competence on this item using the five-point scale, providing insight into their self-awareness of potential validity threats in their assessment practice.

Data collection

The empirical data for this study were drawn from the Chinese English Writing Assessment Competence Corpus (CETWACC; Ke, 2023), a standardized evaluation platform designed to assess the writing assessment competence of early-career EFL teachers in China. CETWACC specifically measures participants' ability to formulate valid, coherent, and pedagogically grounded arguments within the context of EFL writing evaluation—a capacity central to understanding the mediating role of cognitive development in validity argument construction. Higher scores on this assessment reflect greater cognitive and procedural proficiency in writing assessment tasks.To ensure objectivity and reliability in scoring, all participant responses were evaluated by trained raters following a blinded scoring protocol, thereby minimizing potential rater bias. The instrument exhibited excellent inter-rater reliability, with a Cohen's kappa coefficient of 0.92, indicating a high degree of scoring consistency across evaluators. Furthermore, the assessment demonstrated moderate predictive validity when correlated with external indicators of assessment effectiveness, supporting the scale's empirical robustness. Data collection incorporated structured self-report questionnaires alongside supplementary survey instruments, enabling triangulation of responses. All participant submissions were coded according to standardized analytical procedures by raters specializing in language testing and assessment literacy. To verify coding consistency across different testing sites, 40% of the responses were randomly selected for double-coding by independent assessors from separate institutions. This cross-validation process yielded a strong intraclass correlation coefficient (ICC) of 0.84, affirming both the methodological rigor and the reliability of CETWACC as an empirical tool for evaluating writing assessment competence among EFL teachers—particularly in studies examining how cognitive processes mediate the relationship between assessment literacy and validity-oriented scoring practices.

Procedures

This study utilized data collected during the baseline assessment phase of the LTLS Project. Baseline writing assessments were administered within four weeks following participant eligibility screening. The assessment protocol incorporated two core components: (1) teacher-completed

questionnaires, which gathered detailed background information on both the participating teachers and their EFL learners; and (2) standardized writing tasks, which were independently evaluated by trained, blinded raters from the research team to ensure objectivity and minimize scoring bias.

A total of 210 EFL teachers completed a questionnaire-based rubric measuring their perceptions of validity-related outcomes, along with a self-report instrument assessing their own writing assessment literacy. Additionally, all teachers participated in a structured validity argumentation task. Their written responses were evaluated using a standardized scoring framework aligned with contemporary validity constructs. Given the study's primary objective—to examine the mediating role of teachers' cognitive development in their engagement with validity argumentation within writing assessment—the dataset incorporated multiple rater-related variables pertinent to this evaluative process. The final dataset included 210 completed questionnaires and 600 rated writing samples, providing a robust foundation for both quantitative and qualitative analysis.

Data analysis was conducted in two phases. In Phase One, a trained lead coder systematically annotated all semi-structured self-report responses using a predefined coding scheme grounded in assessment literacy theory. In Phase Two, 15% of the coded responses were randomly selected for independent recoding by a second trained analyst to evaluate inter-coder reliability. This two-phase coding procedure enhanced both the consistency and credibility of the qualitative findings, reinforcing the methodological rigor of the analysis.

The significance of mediation analysis in language assessment was further underscored at the European level following the publication of the Companion Volumes by the Council of Europe (2018, 2020). Initially introduced in the Common European Framework of Reference (CEFR) in 2001, mediation was recognized as a core competency to be cultivated within EFL assessment (North & Piccardo, 2016, p. 9). Defined as encompassing activities such as explaining concepts, proposing ideas, and facilitating consensus (North & Piccardo, 2016, p. 29), mediation analysis (Christie, 2002) provides a valuable framework for examining the intermediary cognitive processes that underpin assessment practices—an approach directly aligned with this study's focus on cognitive mediation in validity argument construction.

This study investigated two central sets of relationships: the association between EFL teachers' cognitive characteristics and their assessment literacy, and the link between these cognitive attributes and the quality of their writing assessment practices. To assess the strength and direction of these relationships, we employed Pearson correlation analyses. Furthermore, we conducted mediation analyses using the PROCESS macro for SPSS (Model 4; Hartwell & Aull, 2023) to examine whether teachers' cognitive development in validity argumentation functioned as a mediating variable. According to established methodological guidelines for mediation analysis, a valid mediating effect requires significant correlations between the mediator (cognitive development in validity argumentation) and both the independent variables (teachers' cognitive traits) and the dependent variable (writing assessment quality). These prerequisites served as criteria for establishing the plausibility of mediation effects (Chapelle, 2012; Ke, 2023; Crossley et al., 2025). An alpha level of .05 was adopted for all analyses to determine statistical significance.

To ensure the credibility of qualitative coding, we assessed inter-coder reliability using Cohen's kappa, which yielded a coefficient of 0.78, reflecting substantial agreement between coders. Prior to formal analysis, coders participated in collaborative training using pilot transcripts to align

interpretive standards and resolve discrepancies. This process led to the development of a detailed coding manual containing operational definitions of key categories, annotated examples corresponding to each level of the CETWACC framework, and standardized procedures to promote coding consistency. Throughout the coding phase, the research team held regular calibration meetings to maintain agreement and address emergent challenges. A comprehensive overview of the dataset, coding protocols, and category definitions is provided in Table 1.

Results

Data Screening and Descriptive Statistics. Prior to conducting statistical analyses, the dataset underwent rigorous screening to identify both univariate and multivariate outliers. Following the approach recommended by Arter and McTighe (2001), we identified any data point exceeding 2.2 times the interquartile range (IQR) from the mean as a univariate outlier. A total of 12 such cases were detected. In accordance with Goodman's (2008) guidelines, these values were replaced with the nearest extreme values still within an acceptable range. To identify multivariate outliers, we calculated Mahalanobis distances following the procedure outlined by Hannah et al. (2023). No cases exceeded the critical threshold, indicating no violation of multivariate normality.

After addressing outliers, we examined all variables for normality. Skewness values fell within ±2.0 and kurtosis values within ±9.0, satisfying the normality criteria proposed by Kane (2006) and supporting the use of parametric methods. The final sample comprised 210 EFL teachers with diverse professional backgrounds. Years of teaching experience ranged from 1.5 to 42 ($M = 12.8$, $SD = 8.3$), reflecting substantial variation among participants. Informed by extensive research and theoretical input from the research team, a three-year interval was adopted to meaningfully differentiate levels of teacher competence. Accordingly, participants were divided into seven groups based on teaching experience: Group 1 (< 3 years), Group 2 (3 to < 6 years), Group 3 (6 to < 9 years), Group 4 (9 to < 12 years), Group 5 (12 to < 15 years), Group 6 (15 to < 18 years), and Group 7 (≥ 18 years). Descriptive statistics for all key variables are summarized in Table 1.

Notably, participant responses on the Understanding of Writing Validity scale—designed to measure comprehension of core principles in writing assessment validity—revealed that 52 out of 210 teachers scored below a predefined threshold, suggesting possible gaps in their conceptual understanding. These individuals were identified as having potential "blind spots" in their assessment literacy, particularly regarding the theoretical and practical dimensions of validity in writing assessment.

These results reveal a notable divide within the sample: although many teachers exhibited adequate to strong understanding, a considerable proportion lacked essential competencies required to effectively construct and apply validity arguments. This finding underscores a core motivation for the present study—the urgent need to investigate and support EFL teachers' cognitive development in the area of writing assessment validity.

Table 1 Descriptive statistics

Teaching experience (years)	NO.=210	Understanding of Writing Validity	range	*M*	*SD*
3≤ n≤ 18 years	210	52	25-50	39.74	19.14
Group 1	30	2	25-28	16.26	8.46
Group 2	30	3	27-31	29.33	10.09
Group 3	30	7	29-37	34.09	11.53
Group 4	30	8	32-38	35.64	11.94
Group 5	30	10	34-42	38.30	12.26
Group 6	30	10	36-45	40.75	12.85
Group 7	30	12	30-50	45.37	14.41

Table 2 Correlations between EFL Teachers' Assessment Literacy and Frequency of Argumentation Weaknesses in Validity Arguments

Model	overgeneralization	vagueness	subjectivity	randomness
TWAL cognition	.626	.671	.768*	.654*
TWAL literacy	.769*	.751*	.725*	.374
TWAL emotion	.621	.365	.413	.684*
TWAL consistency	.236	.552	.821*	.512

EFL Teachers' Assessment Literacy and Frequency of Argumentation Weaknesses. Table 2 presents the correlations between teachers' assessment literacy scores and the frequency of weaknesses identified in their validity arguments, specifically across the four dimensions of overgeneralization, vagueness, simplification, and citation deficiency. The analysis revealed that both TWAL cognition and TWAL literacy were significantly associated with rating behaviors indicative of compromised assessment validity. Specifically, higher scores in these domains correlated positively with increased occurrences of overgeneralization (r=.626, .769*), vagueness (r=.671, .751*), subjectivity (r=.768, .725), and randomness (r = .654*, .374). These unexpected positive correlations suggest that greater cognitive and literacy engagement with TWAL may paradoxically correlate with less consistent or less accurate evaluation practices—possibly due to cognitive overload, misinterpretation, or misapplication of assessment principles.

In contrast, TWAL emotion and TWAL consistency showed relatively weak or non-significant associations with rating imprecision, particularly with vagueness and subjectivity. This finding suggests that emotional disposition and self-perceived consistency may have limited direct influence on the clarity or objectivity of scoring behaviors.

Additional findings, summarized in Table 3, examine the role of the CETWACC framework—specifically its dimensions of cultural infiltration and logical thinking—in relation to rating performance. Both constructs demonstrated strong correlations with issues such as overgeneralization and vagueness, as well as with core assessment criteria including content depth, subjectivity, and randomness. While variables such as structural accuracy and linguistic precision

exhibited only moderate associations with rating inaccuracies, a more notable pattern emerged: teachers with weaker logical reasoning skills tended to display more erratic or imprecise evaluative judgments.

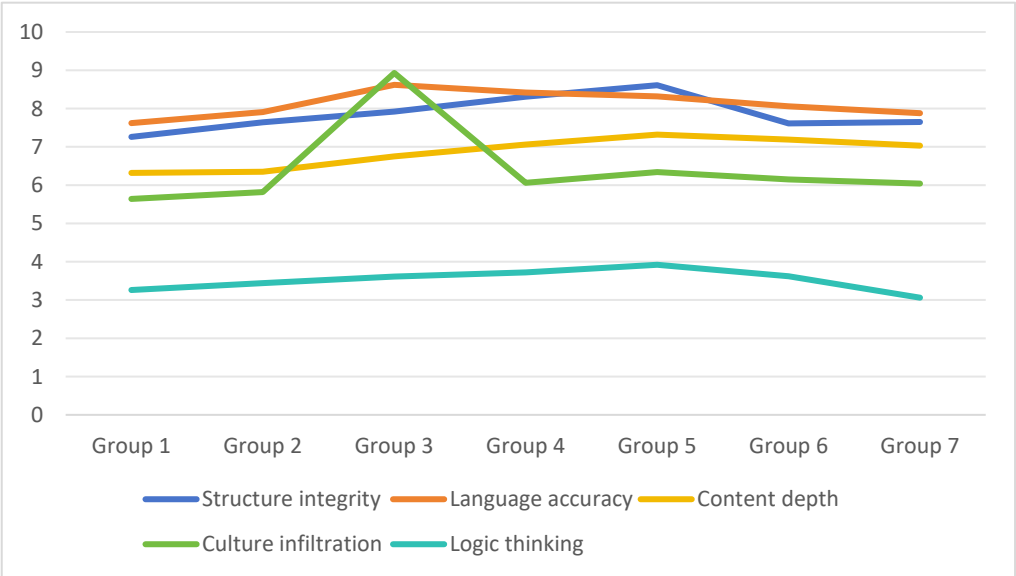

Fig.2 The outcomes scored by EFL teachers' mediating roles of cognitive development in *CETWACC*

EFL teachers' mediating roles of cognitive development in *CETWACC*. Notably, lower logical thinking scores were consistently associated with more pronounced rating inaccuracies, underscoring the essential role of analytical reasoning in upholding the validity and reliability of writing assessment (see Fig. 2). Teachers with underdeveloped logical reasoning skills tended to exhibit subjective, inconsistent, or overgeneralized scoring patterns, thereby undermining the fairness and precision of writing evaluations. Among the seven groups analyzed, Groups 3, 4, and 5 displayed relatively stronger cognitive development in writing assessment, suggesting that moderate teaching experience may support more stable and reflective rating practices. Nevertheless, even these groups underperformed relative to established benchmarks on specific evaluation dimensions—such as structural integrity, language accuracy, and content depth. This indicates that despite advances in general cognitive development, limitations in analytical capabilities may constrain teachers' capacity to deliver nuanced and well-balanced judgments across all scoring criteria. The results reinforce the notion that logical thinking is integral not only to scoring consistency but also to the accurate interpretation and application of assessment rubrics. Consequently, future training initiatives should prioritize the development of teachers' analytical reasoning skills as a core component of efforts to enhance assessment literacy and promote more valid writing evaluation. Similarly, reduced Cultural Infiltration scores were associated with lower rating accuracy when teachers assessed writing from culturally or pedagogically diverse perspectives, with correlation values of r = .617 , .547 , .641 , r=.617,.547,.641, and .631 ∗ .631 ∗ , respectively. Together, these findings highlight the significant impact of both cognitive and cultural factors on the validity and reliability of EFL teachers' assessment practices. They point to a need for targeted professional development that not only strengthens technical assessment skills but also fosters the deeper cognitive and cultural competencies necessary for sound evaluative judgment.

Table 3 Pearson correlations between rating structure variables and rating variables

CETWACC	overgeneralization	vagueness	subjectivity	randomness
Structure integrity	.210	.116	.207	.013
Language accuracy	.428	.334	.281	.127
Content depth	.464	.518	.606*	.416
Culture infiltration	.617*	.547	.641*	.631*
Logic thinking	.776*	.746*	.867*	.817*

Table 4 Pearson correlations between TWAL variables and rating variables

TWAL	Structure integrity	Language accuracy	Content depth	Culture infiltration	Logic thinking
TWAL cognition	.626	.654	.245	.141	.071
TWAL literacy	.641	.767*	.595	.127	.051
TWAL emotion	.665	.621	.474	.402	.105
TWAL consistency	.718*	.784*	.354	.141	.072

TWAL variables and rating variables. The analysis revealed that indicators derived from the TWAL framework were positively correlated with two key dimensions of writing assessment quality: structural integrity (r = .626, .641, .665, .718*) and language accuracy (r = .654, .767*, .621, .784*). These results suggest that higher levels of TWAL-related cognition and literacy are associated with increased evaluative attention to textual organization and linguistic precision, highlighting the potential of TWAL competencies to enhance technical rigor in writing assessment (see Table 4).

In contrast, the TWAL components showed relatively weak correlations with broader cognitive-cultural dimensions such as Cultural Infiltration (r = .141, .127, .402, .141) and Logical Thinking (r = .071, .051, .105, .072). This indicates that, in this context, the TWAL framework does not substantially align with these higher-order constructs, implying that its focus remains more closely tied to assessment-specific knowledge and skills rather than encompassing wider reasoning abilities or cultural awareness.

Furthermore, atypical rating behaviors—those associated with compromised validity according to the CETWACC framework—were significantly correlated with variations in both structural integrity and language accuracy scores. This finding suggests that inconsistencies in evaluating formal textual features often coincide with broader deviations from sound assessment practices.A comprehensive summary of the correlation results is provided in Figure 2. The data indicate a relatively high prevalence of overgeneralization and randomness in scoring, accompanied by notable subjectivity and ambiguity, whereas only a small proportion of teachers consistently applied rubrics in writing assessment. Additionally, Groups 3, 4, and 5 demonstrated comparatively stronger performance, while the remaining groups exhibited relatively weaker outcomes.

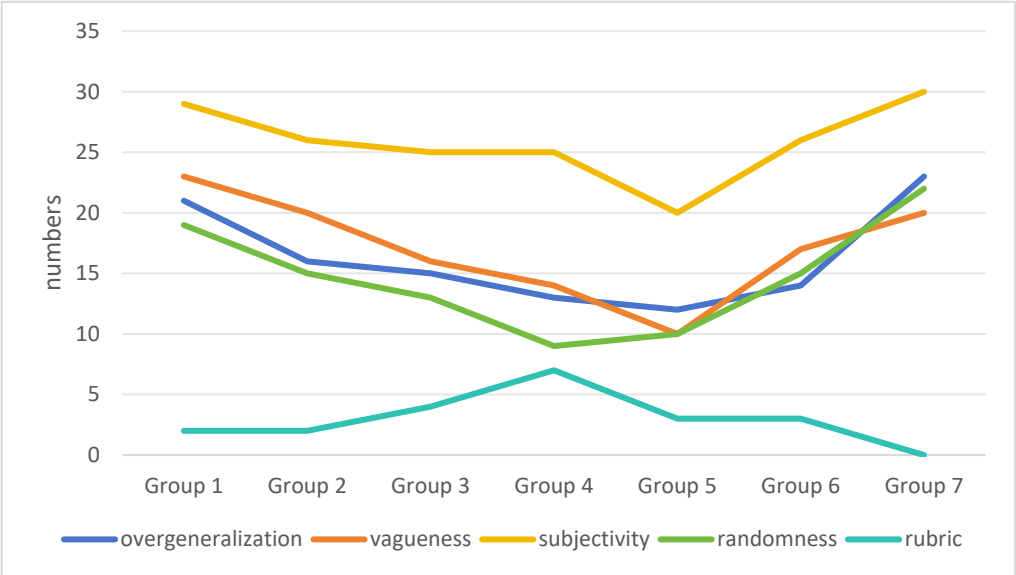

Fig. 3 The differences of different teaching experiences in writing rating standards

A comprehensive summary of all correlation results is presented in Figure 3. The findings indicate a notably high incidence of overgeneralization and randomness in scoring, accompanied by pronounced levels of subjectivity and ambiguity. Furthermore, only a small proportion of teachers consistently applied rubrics in their writing assessments. In terms of group performance, Groups 3, 4, and 5 demonstrated relatively stronger outcomes, while the remaining groups produced assessment results that were comparatively less consistent and accurate.

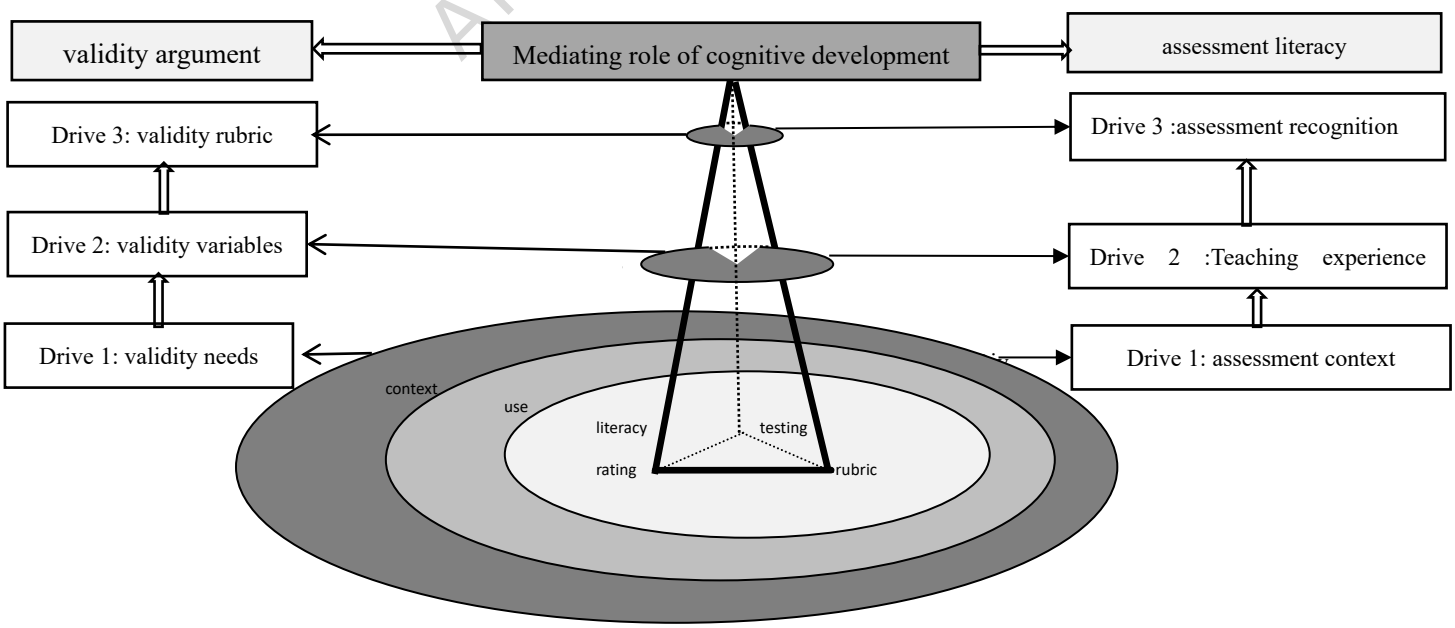

Fig.4 Grounding for the proposed mediating role of cognitive development between validity argument and assessment literacy

Based on the significant correlations observed among the core variables in this study, we developed a conceptual model for constructing validity arguments in writing assessment. Specifically, EFL teachers' cognitive development—as reflected in their performance on writing evaluation tasks—demonstrated strong associations with both the quality of their validity arguments and their professional teaching experience, as measured through the CETWACC framework. These relationships provide empirical support for the proposed validity argument model (see Figure 4), positioning cognitive development as a central mediating mechanism in the construction of validity arguments.

Further reinforcing this model, the robustness of teachers' validity arguments—which serves as an indicator of their cognitive engagement—was significantly correlated with essential components of writing assessment, including test design, rating procedures, and the application of scoring rubrics. These connections highlight the multifaceted nature of validity construction, indicating that cognitive development functions not in isolation but through dynamic interaction with practical assessment activities, thereby shaping how validity arguments are operationalized in authentic classroom contexts.

Given the substantial sample size and the random nature of missing data, it was methodologically sound to integrate both assessment-specific and contextual variables into the model. The proposed framework is built upon three core constructs: teaching experience, assessment literacy, and cognitive development. Within this tripartite structure, cognitive development serves a dual role—as both a driving force and a protective mechanism. Specifically, it sharpens the precision of validity arguments while guarding against conceptual and procedural shortcomings, underscoring its function as the key mediator between teachers' background characteristics and their assessment outcomes.

Discussion

This study examined whether EFL teachers' cognitive development mediates the relationship between their individual characteristics and the quality of validity arguments in writing assessment. We hypothesized that higher levels of TWAL (Teacher Writing Assessment Literacy) traits would enable teachers to conduct more effective writing evaluations, thereby facilitating more accurate and defensible validity judgments. The findings partially supported this hypothesis. Teachers' cognitive mediation—observable in their construction of validity arguments—was indeed associated with distinct and identifiable rating behaviors. Contrary to initial expectations, however, this effect was not driven by writing task characteristics. Instead, it stemmed from internal cognitive processes activated during assessment, particularly teachers' ability to employ validity-oriented reasoning throughout the evaluation process—a finding that directly aligns with the study's focus on the mediating role of cognitive development in validity argumentation.

The following section explores the mechanisms underlying these relationships in greater depth,

further elucidating the mediating interactions among variables across three core domains: TWAL traits, CETWACC indicators, and dimensions of the writing assessment rubrics. Through this tripartite lens, we examine how cognitive mediation shapes the translation of assessment literacy into validity-oriented scoring practices, thereby contributing to a more nuanced understanding of EFL teachers' cognitive development within writing assessment validity arguments.

Mediating roles of Cognitive development variables as potential improvement Between TWAL Characteristics and rating Behaviours. The results of the validity argument analyses offer preliminary evidence for a potential causal model linking TWAL-related traits to rating behaviors within the framework of validity argumentation. Specifically, the findings indicate that EFL teachers may pay insufficient attention to validity principles during writing assessment, leading to systematic scoring inaccuracies. This lack of attentiveness appears to stem from limited engagement in professional development activities designed to promote cognitive growth in constructing and applying validity arguments—a core concern underlying this study's focus on cognitive mediation.

Consequently, such teachers often exhibit disorganized or inconsistent rating patterns—characterized by overgeneralization, vagueness, subjectivity, and randomness—which reflect breakdowns in validity. Furthermore, teachers whose assessments indicated limited familiarity with core validity concepts, such as analytic or criterion-referenced rubrics, also reported stronger negative emotional experiences during scoring. These emotional responses were, in turn, associated with reduced rating accuracy, suggesting that negative affect may function as an additional mediating variable. This affective dimension may signal broader disengagement from or resistance to assessment tasks, potentially influenced by prior educational experiences, cultural attitudes, or deeply held professional beliefs.

The mediating role of negative emotion is particularly significant given previous research establishing that rating literacy strongly predicts rating accuracy (Stevens & Levi, 2005; Chan et al., 2015; McNamara & Knoch, 2019; Liu & Murao, 2025). In this context, the present findings highlight the value of exploring how active engagement with validity arguments enhances assessment quality. Although the results suggest that limited cognitive engagement with validity reasoning may serve as an environmental constraint on rating performance, conclusive evidence regarding causal direction remains preliminary and requires further empirical investigation.

Within the proposed validity argument model, an indirect pathway emerged: EFL teachers' cognitive development influenced rating accuracy through the mediation of rating literacy. The absence of a significant direct effect between cognitive development and rating inaccuracy suggests the presence of broader systemic or contextual factors—such as disciplinary background, institutional culture, or regional assessment practices—that may inhibit the development of rating literacy. These mediating variables could help explain why even teachers with higher cognitive potential sometimes fail to apply validity principles consistently, further underscoring the complex role of cognitive mediation in validity argument construction. While our study offers insights into teachers' cognitive development within a classroom-focused professional development program, the transferability of these findings to purely large-scale, high-stakes validation contexts may be limited. We suggest this as an important direction for future research.

This interpretation aligns with the model's theoretical foundations, which propose that certain teacher characteristics—such as a propensity for subjectivity—are relatively stable and not easily

altered by rating behaviors alone (Ferrara & Qunbar, 2022). Instead, the evidence indicates that cognitive development and rating literacy act as antecedents that shape assessment practices, rather than outcomes determined by them—a distinction that reinforces the mediating role of cognitive development within validity arguments.

To evaluate the robustness of this proposed pathway, we tested an alternative (reversed) model in which inaccurate rating behavior served as the predictor, cognitive development as the mediator, and rating accuracy as the outcome. This reversed model received minimal statistical support, thereby strengthening the original directional hypotheses. Nevertheless, cognitive development as a mediator still accounted for considerable unexplained variance, pointing to the potential influence of additional mediating or moderating variables.

To more conclusively determine the causal mechanisms underlying these relationships, future studies should employ longitudinal research designs grounded in validity argument theory. Such approaches would allow researchers to track the evolution of rating literacy over time and better evaluate its role in enhancing the validity and reliability of writing assessments, ultimately advancing our understanding of how cognitive development mediates the construction of validity arguments in EFL contexts.

Associations between CETWACC and rating outcomes. The findings of this study indicate that while certain CETWACC indicators mediated the relationship between specific EFL teacher characteristics and inaccurate rating behaviors, other rating-related traits appeared to directly influence assessment outcomes without evidence of mediation through cognitive development. In particular, teachers who demonstrated greater cognitive challenges during assessment were more likely to exhibit rating behaviors characterized by overgeneralization, vagueness, subjectivity, and randomness—patterns indicative of weak adherence to validity principles in writing assessment. These observations directly inform the study's central inquiry into how cognitive development mediates validity argument construction.

Interestingly, traits such as structural integrity and language accuracy did not show a significant direct association with any of the identified rating inaccuracies. This observation aligns with prior research. For instance, Ferrara and Qunbar (2022) reported that procedural rating violations were often accompanied by reduced cognitive engagement and limited metacognitive awareness during evaluation. However, the relationship between rating inaccuracy and deeper constructs such as content depth, cultural embeddedness, and logical reasoning remains underexplored in the existing literature.

Previous studies have primarily emphasized the association between formal training in language assessment and assessment literacy, typically measured through benchmark scores (e.g., Jonsson & Svingby, 2013; Karakoç et al., 2025). These works consistently demonstrate that failure to integrate validity arguments into writing assessments is strongly linked to poor rating performance—especially among EFL teachers who display cognitive disengagement, including tendencies toward overgeneralization and arbitrary judgment. More recent studies reinforce this pattern. Knoch and Chapelle (2023), for example, found that negative teacher attitudes toward writing assessment correlated with lower assessment competence, while Ke (2023) identified difficulty in interpreting vague scoring criteria as a barrier to effective rating.

Such assessment deficiencies tend to be more pronounced among teachers with limited formal training, less experience in structured rating contexts, and minimal academic background in

language testing. Kane et al. (2013) also noted that even experienced teachers may revert to intuitive or repetitive scoring behaviors, often relying on subjective impressions rather than rubric-based judgments. The current study's finding—that structural and linguistic dimensions were not significantly associated with rating inaccuracy—further supports the view that many EFL teachers may lack the cognitive readiness or time to engage in deeper evaluative reasoning during assessment tasks (Bachman, 2000; Matta & Hamsho, 2025). These results underscore a broader issue: participating teachers may possess limited awareness of the essential role that validity arguments and rubric use play in ensuring assessment quality—a limitation that directly implicates the mediating function of cognitive development. Such limitations are likely exacerbated in high-stakes testing environments, where precision, objectivity, and consistency are critical.

Regarding the relationship between CETWACC dimensions and rating behaviors, this study found that difficulties in areas such as content depth, cultural embeddedness, and logical reasoning were significantly associated with increased rating inaccuracy among EFL teachers. This finding contrasts with that of Kim (2020), who reported no meaningful association between the structural aspects of writing ratings and the quality of validity arguments. Moreover, Kim observed that ESL teachers with stronger language structure traits paradoxically demonstrated weaker rating performance. These discrepancies may stem from differences in sample characteristics: Kim's study involved a smaller cohort of ESL teachers, whereas the present research examined a larger sample of EFL instructors, many of whom had limited training or exposure to language testing. Such variations highlight considerable methodological diversity—particularly in participant profiles and measurement tools—across studies in this domain.

To our knowledge, this study is the first to explicitly investigate the potential link between EFL teachers' cognitive mediation and inaccurate rating behaviors within a validity argument framework for writing assessment. As such, these findings provide new insights into how cognitive and assessment-related traits interact to shape the reliability and validity of teacher-led writing evaluation, positioning cognitive development as a key mediating mechanism in the construction of validity arguments.

Associations between TWAL variables and rating variables. An in-depth analysis of the validity argument model revealed a paradoxical pattern: higher scores on certain TWAL indicators were associated with more frequent negative rating behaviors among EFL teachers during writing assessment tasks. Specifically, teachers who demonstrated lower cognitive engagement with validity principles were more likely to exhibit subjective, vague, overgeneralized, and inconsistent rating patterns. This suggests that although TWAL indicators may reflect a general awareness of assessment concepts, without a firm grounding in validity reasoning, such awareness does not necessarily translate into more accurate or consistent scoring—a finding that directly speaks to the mediating role of cognitive development in validity argumentation.

Beyond the internal dynamics of the validity argument model, the study also identified an unexpected relationship: greater perceived difficulty in rating was directly linked to lower rating accuracy. This finding challenges conventional assumptions that inaccuracies arise solely from technical deficiencies in assessment knowledge. Instead, it aligns with prior claims—such as those by Knoch (2009)—that affective and psychological factors, including emotional disengagement or low assessment self-efficacy, may significantly influence rating performance. From this perspective, the connection between cognitive development and the formulation of validity arguments appears

to reflect a complex interaction between personal dispositions and assessment behaviors, raising questions about the capacity of conventional rating rubrics to account for these underlying emotional and motivational dimensions.

Additional patterns emerged across both observational (TWAL) and self-reported (CETWACC) data. Interestingly, teachers who reported higher perceived competence in constructing validity arguments were sometimes more prone to rating inaccuracies. This counterintuitive association resonates with findings by Fulcher (2010), who noted that strong confidence in assessment reasoning does not always translate into consistent rating performance—particularly in the absence of structured supports such as well-defined rubrics (see also McNamara & Knoch, 2019). Importantly, this should not be interpreted as evidence that cognitive development causes poor rating. Rather, it underscores structural and contextual barriers that hinder the effective application of assessment knowledge, reinforcing the need to examine mediating mechanisms within validity argument frameworks.

As Kunnan (2000) argued, the root cause of assessment errors often lies not in cognitive limitations but in the lack of training in constructing and applying validity arguments. Many EFL teachers have limited exposure to language assessment scholarship, which restricts their access to best practices and tools necessary for valid rating. The results thus point to a systemic issue: writing assessment in EFL contexts is often outcome-oriented and inadequately aligned with process-based models of professional learning. This misalignment impedes the development of cognitive competence and contributes to superficial or inconsistent assessment practices—underscoring the centrality of cognitive mediation in bridging assessment literacy and valid scoring.

Furthermore, teachers with higher rates of rating inaccuracies tended to report more negative emotional responses to assessment tasks and exhibited lower motivation for self-directed professional development. These individuals were also more likely to have received little or no formal training in language testing, reinforcing the connection between inadequate preparation and diminished assessment quality. Together, these findings suggest that contextual factors—such as training quality, rubric clarity, and emotional engagement—mediate the effectiveness of broader educational reforms in language assessment, operating alongside cognitive development to shape validity argument construction.

The implications are substantial. They call for targeted, multidimensional interventions that not only enhance cognitive and technical assessment literacy but also promote emotional resilience and sustained professional growth. Future research should adopt longitudinal approaches to examine how cognitive, affective, and contextual variables interact over time to shape rating practices. Such work is essential for designing teacher education programs that cultivate both sharper validity reasoning and greater judgmental stability, with explicit attention to fostering cognitive mediation within validity arguments.

Ultimately, fostering the cognitive development of EFL teachers within supportive, well-structured assessment environments is crucial for advancing fairness, accuracy, and defensibility in writing assessment. This approach moves beyond isolated technical skills toward a more holistic model of professional competence—one rooted in both cognitive rigor and reflective practice, and centrally organized around the mediating role of validity assessment in teachers' cognitive development.

Limitations

This study represents a pioneering effort to examine the interrelationships among three fundamental dimensions of EFL writing assessment: teacher characteristics, teacher attitudes, and rating outcomes. As an initial exploratory investigation, no statistical corrections for multiple comparisons were applied in the analysis of associations across these domains. Therefore, the findings should be interpreted with caution and regarded as a preliminary framework to be validated through future confirmatory research. Replication with larger and more diverse samples will be essential to substantiate and extend these early insights, particularly in relation to the mediating role of cognitive development in validity argument construction.

Due to the cross-sectional design of the study, causal inferences regarding the relationships between EFL teachers' attributes, their mediation of rating behaviors, and final assessment outcomes remain provisional. The research adopted a mixed-methods approach, integrating an objective observational tool (TWAL) and a teacher self-report instrument (CETWACC) to evaluate inaccuracies in rating behaviors. Notably, both measures showed significant associations with overall rating quality, indicating a mediating convergence between teachers' self-perceptions and their externally observed assessment practices. This alignment strengthens the credibility of the results and highlights the practical utility of validity-centered rating frameworks for understanding cognitive mediation.

The findings also suggest that EFL teachers with lower levels of rating competence—particularly those exhibiting consistent validity-related inaccuracies—may benefit from targeted professional development in assessment literacy. However, structural constraints within test-oriented educational systems may inhibit such growth. A lack of opportunities for reflective learning and constructive feedback can obstruct the development of accurate rating skills and reinforce flawed assessment practices, thereby impeding the cognitive mediation essential for robust validity arguments. Previous research has highlighted the value of validity argument frameworks for tracing causal pathways in educational assessment, particularly within longitudinal designs (Cronbach, 1988; Bachman & Palmer, 1996; Taylor et al., 2023; Ferrara & Qunbar, 2022; Xu & Zheng, 2025). While this study enhances our understanding of the correlational patterns among rating-related variables, its cross-sectional nature precludes insights into how validity reasoning and cognitive mediation develop over time (Correnti et al., 2022).

To address this limitation, future studies should employ longitudinal approaches to investigate how teacher cognition, attitudes, and assessment behaviors evolve across different stages of professional experience. Such research is vital for elucidating the mechanisms through which cognitive and affective factors influence rating performance, particularly the mediating pathways central to validity argumentation. In turn, these insights will inform the design of sustained, evidence-based interventions aimed at improving both the accuracy and validity of EFL writing assessments. Ultimately, a deeper understanding of these developmental dynamics is essential for fostering a more reflective, equitable, and professionally grounded assessment culture—one in which teachers' cognitive development serves as a key mediator in the construction of meaningful validity arguments for writing assessment.

Conclusion

This study demonstrates that EFL teachers' writing assessment practices serve as a critical mediating factor between teacher-related characteristics and the occurrence of inaccurate rating behaviors. By identifying potential causal pathways linking individual teacher attributes to specific assessment outcomes, the research contributes meaningful insights to the expanding literature on writing assessment. A key contribution of this study is its emphasis on a crucial yet often overlooked issue in teacher education: the persistent gap between EFL teachers' theoretical understanding and their practical application of validity arguments during the rating process—a gap that directly implicates the mediating role of cognitive development. Despite its significance, this disconnect is frequently neglected in both professional development programs and formal assessment training, undermining the validity, fairness, and credibility of writing evaluations.

If confirmed through longitudinal research, these findings could have important implications for designing targeted interventions aimed at enhancing assessment literacy among EFL teachers. Crucially, such initiatives should aim not only to improve rating accuracy but also to foster what might be termed an "optimal rating state"—characterized by greater validity, reliability, and consistency in assessment practices. Although a consensus definition of this state has yet to be established, existing research underscores the mediating function of well-structured scoring rubrics in facilitating valid and consistent judgment (Correnti et al., 2022; Stephens & Sarkar, 2025), further highlighting the interplay between cognitive development and assessment tools within validity arguments.

Consistent with this view, the present study underscores the importance of integrating cognitive development strategies into teacher training programs. This includes fostering a deeper conceptual grasp of validity principles alongside the practical skills needed to apply them in actual rating situations. Equipping teachers with the cognitive tools for rigorous validity reasoning will not only enhance individual assessment performance but also promote more equitable and defensible evaluation practices throughout educational systems—thereby strengthening the mediating mechanisms that link assessment literacy to valid writing assessment outcomes.

Ultimately, these findings advocate for a systemic reconceptualization of how writing assessment is understood and implemented in EFL contexts. Strengthening teachers' assessment literacy has the potential to elevate the overall integrity and effectiveness of language testing. Future research should continue to investigate how sustained cognitive and pedagogical support can facilitate this transformation over time, with particular attention to the mediating role of validity assessment in EFL teachers' cognitive development within writing assessment validity arguments.

Declarations

Consent to participate Informed consent was obtained from all the participants involved in this study.

Ethical approval Ethical approval for this study was granted by the Human Research Ethics Committee of Taizhou University (Approval Number: 2025L-SC-011). The research was conducted in strict accordance with the ethical principles outlined in the Declaration of Helsinki, ensuring the protection of participants' rights and welfare throughout the research process. All procedures involving human participants adhered to the approved guidelines and institutional requirements.

Data Availability Statement The raw data supporting the conclusions of this article will be made available by the authors without undue reservation. All research data can be openly shared and are freely available in the Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TDIF2E)

Conflict of interest The authors have no potential conflicts of interest to disclose.

References

Arter J, McTighe J (2001) Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Corwin Press.

Bachman L F (2005) Building and supporting a case for test use. Language Assessment Quarterly, (1):1-34.

Bachman L F, Palmer A S (1996) Language testing in practice: Designing and developing useful language tests (Vol. 1). Oxford University Press.

Bachman, L F (2000). Modern language testing at the turn of the century: Assuring that what we count counts. Language testing, 17(1), 1-42.

Borg S. (2003).Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. Language Teaching. 36(2):81-109. doi:10.1017/S0261444803001903

Chan S, Inoue C, Taylor L (2015) Developing rubrics to assess the reading-into- writing skills: A case study. Assessing Writing, (26): 20-37.

Chapelle C A (2012) Validity argument for language assessment: The framework is simple…. Language Testing, 29(1), 19-27.

Choi S, McMaster K L, Kim N(2025) Toward the fair and valid use of curriculum-based measurement for students with intensive writing needs and linguistically diverse backgrounds. Assessing Writing, 65, 100948.

Christie, Frances. (2002). Classroom discourse analysis: A functional perspective. London: Continuum.

Correnti R, Matsumura L C, Wang E L, Litman D, Zhang H (2022) Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. Computers and Education Open, 3, 100084.

Council of Europe (2018) Common European Framework of Reference for Languages: Learning, teaching, assessment companion volume with new descriptors. Strasbourg: Council of Europe.

Council of Europe (2020) Common European Framework of Reference for Languages: Learning, teaching, assessment–companion volume. Strasbourg: Council of Europe Publishing.

Cronbach L J (1988) Five perspectives on validity argument[A]. In NJ: Lawrence Erlbaum,1988: 3-17.

Crossley S A, Kim M, Wan Q, Allen L K, Tywoniw R, McNamara D (2025) Assessing writing quality using crowdsourced non-expert comparative judgement ratings. Assessment in

Education: Principles, Policy & Practice, 32(1), 33-59.
Dorsey D W, Michaels H R (2022) Validity arguments meet artificial intelligence in innovative educational assessment. Journal of Educational Measurement, 59(3), 267-271.
Ferrara S, Qunbar S (2022) Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration.Brain Behaviour. 5, 22 – 44.
Folger T D, Bostic J, Krupa E E (2023) Defining test-score interpretation, use, and claims: Delphi study for the validity argument. Educational Measurement: Issues and Practice, 42(3), 22-38.
Fulcher G (2010) Practical Language Testing. London, UK: Hodder Education, 2010.
Goodman N D, Tenenbaum J B, Feldman J, Grifiths T L (2008) A rational analysis of rule-based concept learning. Cognitive Science. 32, 108-154.
Hannah et al (2023)Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits account. Comput. Brain Behaviour. 5, 22-44.
Hartwell K, Aull L (2023) Editorial Introduction – AI, corpora, and future directions for writing assessment. Assessing Writing, 57, 100769.
Jonsson A, and Svingby G (2007) The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review, 2(2), 130-144.
Kane M T (2006)Validation[A]. In R. Brennan(ed.). Educational Measurement [C]. Westport, CT: Greenwood Publishing,17-64.
Kane M T (2013) Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.
Karakoç A I, Gu P, Ruegg R (2025) Designing a rating scale for an integrated reading-writing test: A needs-oriented approach. Assessing Writing, 64, 100918.
Ke Y (2024) Examining simultaneous pausing on the cognitive writing process: a micro-formative writing assessment. Current Psychology, 1-12.
Ke Y (2023) Integrated positive and negative analyses of cognitive-mediation strategies in the social quarrels. Humanities and Social Sciences Communications, 10(1), 1-8.
Kim H. (2020). Effects of rating criteria order on the halo effect in EFL writing assessment: a many-facet Rasch measurement analysis. Language Testing in Asia, 10(1), 16.
Knoch U (2009) Diagnostic Writing Assessment: The Development and Validation of a Rating Scale. Peter Lang:Flevelurt.
Knoch, Chapelle (2018) Knoch,U. and C. A. Chapelle. Validation of rating processes within an argument-based framework. Language Testing, (4): 477-499
Kunnan A J, Jang E E (2009) Diagnostic feedback in language assessment[A]. In M. H. Long and C. J. Doughty(eds.). The Handbook of Language Teaching [C]. Malden, MA: Wiley Blackwell, 610-627.
Kunnan A J. (Ed.) (2000) Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (Vol. 9). Cambridge University Press.
Liu K, Murao R (2025) Reliability and validity assessment of working memory measurements. Applied Psycholinguistics, 46, e3.
Matta M, Hamsho N (2025) Consequences of Response Formats on Racial and Ethnic Bias and Fairness in Writing Assessments. School Psychology Review, 1-14.
McNamara T, Knoch U (2019) Fairness, justice and language assessment. Oxford University Press.

Nadal Sanchis, L., & Bello Viruega, I (2024) Insights from an empirical study on communicative functions and L1 use during conceptual mediation in L2 peer interaction. International Review of Applied Linguistics in Language Teaching, 62(4), 2149-2171.

Nadal, L & Sarah T (2021) Mediation and German as a foreign language in non-contexts Immersion: An Experimental Analysis. Theoretical and Applied Linguistics (RLA) 59. 111–132.

North, Brian & Enrica Piccardo (2016) Developing illustrative descriptors of aspects of mediation for the CEFR:A Council of Europe project. Strasbourg: Council of Europe.

Piantadosi S T, Tenenbaum J B, Goodman N D (2016) The logical primitives of thought: empirical foundations for compositional cognitive models. Psychology Review,123, 392 – 424 .

Romig J E, Olsen A A, Medina E, Tulloh A (2025) Criterion validity evidence and alternate form reliability of curriculum-based measures of written expression for eighth grade students. Assessing Writing, 66, 100958.

Sawaki Y (2024) Validity Argument in Language Testing: Case Studies of Validation Research. Language Testing, 41 (1) , pp.227-229.

Sireci S G (2013) Agreeing on validity arguments. Journal of Educational Measurement, 50(1), 99-104.

Stephens G C, Sarkar M (2025) Valid concerns: Considerations for reviewing manuscripts with validity arguments. Anatomical Sciences Education.

Stevens D D, Levi A (2005) leveling the field: Using Rubrics to achieve greater equity in teaching and grading. Essays on Teaching Excellence.

Taylor L (2023) Reframing the discourse and rhetoric of language testing and assessment for the public square. Language Testing, 40(1), 47-53.

Thwaites P, Jadoulle P, Paquot M (2025) Comparative judgment in L2 writing assessment: Reliability and validity across crowdsourced, community-driven, and trained rater groups of judges. Assessing Writing, 65, 100937.

Thwaites P, Paquot M (2025)Testing crowdsourcing as a means of recruitment for the comparative judgement of L2 argumentative essays. Journal of Second Language Writing, 68, 101207.

Toulmin S E (2003)The Uses of Argument (updated edn.). Cambridge: Cambridge University Press.

Toulmin S E (1958)The Uses of Argument. Cambridge: Cambridge University Press

Wu Q (2025). Comparative Judgment: Building a Shared Consensus Over Rater Variation in Assessing Second Language Writing Performance. SAGE Open, 15(2), 21582440251346346.

Xu J, Zheng Y (2025) Does student assessment literacy matter between motivational constructs and engagement in L2 writing? A survey of Chinese EFL undergraduates. Assessing Writing, 64, 100916.