Language Testing in Asia

RESEARCH Open Access

Rater cognitive processes in integrated writing tasks: from the perspective of problem-solving

Wenfeng Jia[1] and Peixin Zhang[2*]

*Correspondence:
zhangpeixin@xmu.edu.cn

[1] School of Translation Studies, Shandong University, 180 Wenhuxi Road, Weihai, Shandong 264209, China
[2] College of Foreign Languages and Cultures, Xiamen University, 422 Siming South Road, Siming, Xiamen 361005, China

Abstract

It is widely believed that raters' cognition is an important aspect of writing assessment, as it has both logical and temporal priority over scores. Based on a critical review of previous research in this area, it is found that raters' cognition can be boiled to two fundamental issues: building text images and strategies for articulating scores. Compared to the scoring contexts of previous research, the TEM 8 integrated writing task scoring scale has unique features. It is urgent to know how raters build text images and how they articulate scores for text images in the specific context of rating TEM8 compositions. In order to answer these questions, the present study conducted qualitative research by considering raters as problem solvers in the light of problem-solving theory. Hence, 6 highly experienced raters were asked to verbalize their thoughts simultaneously while rating TEM 8 essays, supplemented by a retrospective interview. Analyzing the collected protocols, we found that with regard to research question 1, the raters went through two stages by setting building text images as isolated nodes and building holistic text images for each dimension as two sub-goals, respectively. In order to achieve the first sub-goal, raters used strategies such as single foci evaluating, diagnosing, and comparing; for the second sub-goal, they mainly used synthesizing and comparing. Regarding the second question, the results showed that they resorted to two groups of strategies: demarcating boundaries between scores within a dimension and discriminating between dimensions, each group consisting of more specific processes. Each of the extracted processes was defined clearly and their relationships were delineated, on the basis of which a new working model of the rating process was finalized. Overall, the present study deepens our understanding of rating processes and provides evidence for the scoring validity of the TEM 8 integrated writing test. It also provides implications for rating practice, such as the need for the distinction between two types of analytical rating scales.

Keywords: Writing assessment, Rating processes, TEM 8 rating context, Problem-solvers, Building text images, Strategies for articulating scores

Introduction

The purpose of language testing is to use scores to infer the underlying language ability of candidates (McNamara, 1996; Bachman & Palmer, 2010). For most high-stakes and large-scale writing assessments, it is still a common practice to employ human raters to

produce scores. Under these circumstances, scores are resultant of interaction between raters, rating scales, and compositions (Green, 1998; Weigle, 2002), and the raters' cognitive process is the crux of the matter. Rater cognition has both logical and temporal priority over scores, and without a thorough and detailed description of the raters' scoring process, it is impossible to argue for the scoring validity of a writing test (Heidari et al., 2022). According to Knoch and Chapelle (2018), one warrant for backing the scoring validity is the alignment between the raters' cognitive processes and the writing construct to be tested in writing tasks.

However, there is usually no definitive procedure for rating compositions. DeRemer (1998) claims that rating is a process of problem-solving. Typically, due to the vagueness of the rating scale, the problem of scoring is considered to be "ill-structured," and therefore, raters have to come up with creative solutions for this problem. In one article exploring what the rating criteria really meant to the raters, Lumley (2002) notes that "[t]he rules and the scale do not cover all eventualities, forcing the raters to develop various strategies to help them cope with problematic aspects of the rating process." In this way, raters are better identified as problem solvers in the process of rating writing scripts.

In the recent two decades, integrated writing tasks have increasingly been adopted in both large-scale language test and classroom-based writing assessments (Knoch & Sitajalabhorn, 2013; Cumming, 2014). According to Knoch and Sitajalabhorn (2013), a writing task can only be called integrative if three requirements are met. First, the task input must contain materials of written texts. Writing tasks with only pictures as input rather than words are not integrative writing tasks. Second, the writing prompts must clearly indicate how the source materials are to be integrated into the writing. Third, the rating scale must reflect the integration requirements. Following this trend, the writing module of TEM8 (Test for English Majors, Band 8), a nationally standardized English proficiency test in mainland China, also replaced the independent writing task with a writing task based on reading in its latest reform in 2016, which meets the three requirements above (see Supplementary I). Since its first administration, it has been "welcomed by teachers and students in various colleges for its authenticity" (Liu & Fan, 2020). Surprisingly, however, little research has been conducted to investigate raters' cognitive processes on this new type of writing task.

Compared with other rating scales widely used in language testing, the TEM8 rating scale has three distinctive features (see Supplementary II). Firstly, the scoring criteria or dimensions are quite broad, with each criterion containing a large number of descriptors. The three criteria are content, organization, and language use, and the first dimension contains more than 10 descriptors. This feature makes the TEM8 rating scale rather "thick," which is in contrast to the rating scale whose criterion includes only one descriptor, such as the scale of STEP in Australia (Special Test of English Proficiency) in which each criterion includes only one descriptor (Lumley, 2005), or the scale of STAP (Spanish Test for Academic Purposes) in which each criterion corresponds to a very specific construct, such as cohesion, grammatical accuracy, and so on (Mendoza & Knoch, 2018). Secondly, the score distribution of the rating scale of TEM8 is uneven, with dimensions of content, organization, and language use receiving 10, 3, and 7 points, respectively, which is in contrast to most other scales with the same scores for each dimension, for example, in the ESL Composition Profile (Jacobs et al., 1981). Thirdly, the rating scale

of TEM8 juxtaposes the descriptors for assessing both summary and argumentation in three dimensions as one requirement of an integrated task. In a word, the rating scale of TEM8 is quite unique and we still do not know how raters cognitively use this type of scale to produce scores.

To address this gap, the present study will focus on raters' cognitive process in the TEM8 integrated writing rating environment from the problem-solving perspective, i.e., how raters cognitively solve the problem of rating. This study is expected to advance our knowledge of raters' cognitive processes, particularly in relation to integrated writing tasks with an analytic rating scale. In addition, the findings will contribute to our understanding of rating validity for the TEM8 integrated writing task.

Literature review

Problem-solving in cognitive psychology

In cognitive psychology, Frensch and Funke (1995) contend that a problem is not defined by the task itself, but by "the interaction between task characteristics and person characteristics" (p.28). In this sense, a problem only exists if there is a distance between the task situation and the solvers, and Ormrod (2012) emphasizes that problem-solving involves "deliberate and controlled mental processes" (p. 402). Similarly, Anderson (2015) defines problem-solving as "goal-directed behavior that often involves setting sub-goals to enable the application of operators" (p.183). Specifically, the problem-solving behavior is clearly organized toward an overarching goal. However, the problem is not solved in one fell swoop; instead, it is decomposed into sub-goals of different states, which is a representation of the problem in degree of solution. With the help of the operators, mainly acquired by discovery or by direct instruction, one problem state is transformed into the next problem state until the whole problem is solved. Although in each state, there are many ways the problem solver can choose to change the state, and the problem solver tends to adhere to the principle of difference-reduction. As stated by Anderson (2015), problem solvers are defined as "choosing operators that transform the current state into a new state that reduces differences and resembles the goal state more closely than the current state" (p.192). According to Ormrod (2012), apart from these procedures, problem-solving also involves a step of looking back, i.e., evaluating the overall effectiveness of problem-solving efforts in order to learn some lessons for possible future use.

Existing models of rater cognitive processes

As stated by DeRemer (1998) and Lumley (2002) above, there is no clear procedure for raters to follow when they are rating compositions, and thus, they are best regarded as problem solvers, who rely mainly on themselves for solutions. In the research field of the rating process of raters, Freedman and Calfee (1983) were among the first who noticed that raters can be regarded as problem solvers and they put forward the first model that symbolized the rating process. In applied linguistics, models consist of definitions of categories or processes and their relationships (Flower & Hayes, 1981) that are important for understanding cognitive activity. Building process models based on inferring the cognitive process from the collected data is also essential for researching problem-solving activity (Kluwe, 1995). In the model of Freedman and Calfee (1983), three stages of processes were identified as crucial in rating a composition: (1) reading and comprehending

text to create a text image, (2) evaluating a text image and storing impressions, and (3) articulating evaluation. Text image is defined as the mental representation of essays, which is the prerequisite for scoring. Subsequently, Cumming et al. (2002), Wolfe (2005), and Lumley (2005) also constructed models as simple and symbolic representations of rating procedure. In the following, we will outline the three models mentioned above, followed by a critical commentary on both the models and other recent studies, until we come to the two research questions of the present study.

The model of Cumming et al. (2002)

The model of Cumming et al. (2002) consists of the proto-typical decision sequence for scoring TOEFL essays and the descriptive framework for decision-making, represented by the coding schema in Table 1. The rater went through three stages for rating: scanning the composition for surface, engaging in interpretation strategies, and articulating a scoring decision. Table 1 further instantiates the cognitive stages by listing both what raters experienced cognitively and the textual features to which they attended. All cognitive processes in Table 1 can be divided into two broad categories: interpretation strategies and judgment strategies. Interpretation strategies consist of reading strategies aimed at understanding the essay, whereas judgment strategies are evaluation strategies aimed at formulating a rating or score.

This model had a strong influence on other related research. In particular, the coding scheme for evaluation processes in Table 1 became the main source for coding verbal protocols in subsequent research, for example, Barkaoui (2007, 2010), Li and He (2015), and Heidari et al. (2022), which aimed at comparing the rating processes across different rating contexts, mainly using a holistic and analytic rating scale. Their general conclusion was that rater cognitive behaviors are flexible and malleable, being subject to the specific requirement of rating scales. Relatedly, Cumming (1990) himself used a similar

Table 1 The coding schema for the rating process in the TOEFL rating context

Self-monitoring focus	Rhetorical and ideational focus	Language focus
Interpretation strategies		
Read or interpret writing prompts	Interpret unclear expressions	Observe layout
Read or reread compositions	Discern rhetorical structures	Classy errors into types
Envision the personal situation of the authors	Summarize ideas or propositions	Edit phrases for interpretation
Scan whole compositions		
Judgment strategies		
Decide strategies for reading	Assess reasoning, logic, and development	Assess quantity or overall production
Consider personal response	Assess task completion	Assess comprehensibility
Define or revise your own criteria	Assess relevance	Consider the gravity of errors
Compare with other composition	Assess coherence	Consider error frequency
Summarize or tally judgment collectively	Assess originality or creativity	Assess fluency
Articulate general impression	Assess redundancies	Consider lexis
	Assess text organization	Consider syntax or morphology
	Assess style, register, or genre Rate ideas or rhetoric	Consider spelling or punctuation Rate language overall

coding scheme to compare the rating processes of expert and novice raters and found that, in general, experts used a wider range of cognitive processes and more frequently than novices.

Wolfe's model (2005)

Wolfe's model is based on his generalizations from the analysis of raters' verbal protocols, as shown in Fig. 1. His model emphasized the interaction between the text image and the rating process. According to it, raters read texts written by students and formed mental images of the text. Of course, the text images formed might differ from rater to rater because of their different experiences. Once the text image was created, evaluation (including monitoring, reviewing, and deciding) and justification (including diagnosis, rationale, and comparison) were functioning for articulating scores.

Similar to Cumming et al.'s (2002) model, Wolfe's model was also contextualized using a holistic rating scale. The merit of this model lies in a clearer explanation of the role of building text images and its relationship to other rating processes. Related to this model, Wolfe et al. (1998) demonstrated that groups of raters with different levels of expertise differed cognitively in the following three aspects: raters with high expertise tended to cite more general features of compositions, to use more of the language provided by the test developer in the descriptors, and to use a top-down approach to essay scoring; whereas raters with low expertise tended to focus more on specific features, to use more self-generated descriptive words not found in the rubric, and to use a bottom-up approach to essay scoring.

Lumley's model

In contrast to the two models above, Lumley's (2005) model was contextualized by the use of an analytic rating scale in the Australian test of STEP. His model was also derived from verbal protocol analysis, as shown in Fig. 2. According to Fig. 2, raters went through three stages before producing a score, namely initial reading, scoring, and summarizing. It has been mentioned in the "Introduction" section that the STEP scale is rather "thin"

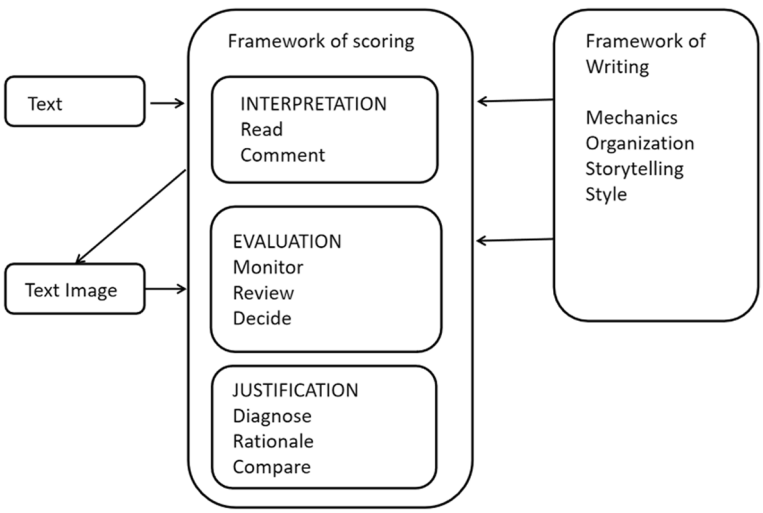

Fig. 1 Wolfe's model of rating process (2005)

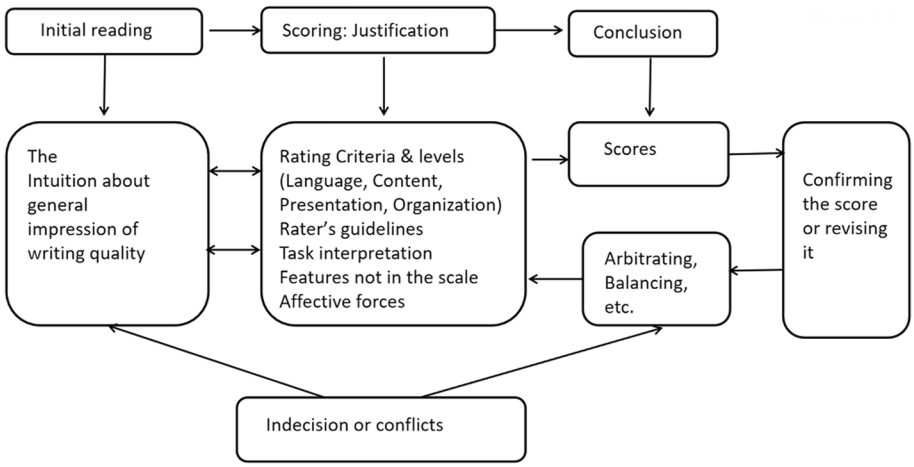

Fig. 2 Lumley' model of the rating process (abridged)

in that each rating dimension contains only one descriptor. Closely related to this, raters in Lumley's study were expected to map the quality of writing in a more rigid way with the descriptor, in order to seek "transparency" in the scores. Under these circumstances, the text image was considered to be subjective and unfavorable. For this reason, a text image is not explicitly shown in this model.

In a more general perspective, it is found that rater cognition can be boiled down to two fundamental issues: building text images and strategies for articulating scores. Building text images is regarded as a prerequisite for scoring (Freedman & Calfee, 1983; Cumming et al., 2002) and is most clearly expressed in Wolfe's model (2005) as above. As to the strategies for articulating scores, these models are also informative. For example, scoring strategies such as balancing and arbitrating are listed in Lumley's model (Fig. 2). Based on these two core issues, we can find some parallel correspondence between the above models. Specifically, the stage of interpretation strategies in Cumming et al.'s model, the interpretation stage in Wolfe's model (see Fig. 1), and reading for general impression in Lumley's model (see Fig. 2) all embody how raters build text image in rating, and they are counterparts to each other. On the other hand, the stage of articulating strategies in Cumming et al.'s model, the Justification in Wolfe's model (see Fig. 1), and the Justification and Conclusion in Lumley's model (see Fig. 2) are all embodiments of how raters articulating scores and they are counterparts of each other. In a nutshell, although different models used different names for describing processes, the two shared core issues are building text images and strategies for articulating scores.

In addition to the three models, there are two more recent studies tapping on the cognitive processes, but they do not construct models. Zhang (2016) studied the raters' cognitive process in the context of CET 4 rating and extracted about 10 categories of processes, including comparison, diagnosis, and monitoring. Yan and Chuang (2023) extracted 17 categories of cognitive processes, such as commenting on the thesis statement, commenting on the use of sources, and commenting on the severity of errors. These three classic models and related recent studies have contributed to our general understanding of raters' cognitive processes in variant contexts of rating compositions.

However, we still can identify research gaps when the following two aspects are considered.

Firstly, it is obvious that all the above three models and recent studies are aimed at a different rating context, and for a new rating context such as rating TEM8 integrated compositions, it is necessary to construct a new working model. On the one hand, the models of Cumming et al. (2002) and Wolfe's (2005) were designed to explain the rating processes of using a holistic rating scale, and therefore, they are not suitable for the TEM8 situation with an analytical rating scale. For example, the coding schema as shown in Table 1 that was widely followed by other research is not fit for the TEM8 writing task as one integrated writing task. On the other hand, although Lumley's (2005) model is indeed based on the analytic scale of the STEP writing test, the form of this analytic scale is quite different from that of TEM8, as mentioned in the "Introduction" section. Neither, Lumley's (2005) models is far from suitable for the TEM8 assessment context. The studies of Zhang (2016) and Yan and Chuang (2023) were based on CET4 in China and a placement test in an American university, using a holistic rubric and a holistic profile-based rating scale respectively. Again, their rating process taxonomies are not suitable to explain that of TEM8.

Secondly, from the perspective of problem-solving theory, the expressions for the above models are not without room for improvement. Clearly, viewing the rating process as a problem-solving activity means that rater cognitive behavior is characterized by the general features of problem-solving activity, such as decomposing the task into sub-tasks, setting sub-goals, and choosing operators to shorten the distance between the status quo and the final state, as elaborated by Ormrod (2012) and Anderson (2015) above. However, it seems that certain concrete rating processes in the above models are not arranged in a strict chronological order. For example, in Lumley's (2005) model as shown in Fig. 2, although evaluation and justification as a whole can be regarded to follow the stage of interpretation, the specific sub-processes such as monitoring, reviewing, and justifying were arranged in such a way without considering their order. As a result, the readers are unclear about their chronological relationship. This is not consistent with the problem-solving theory, in which one problem will enter into a new stage after the sub-goal of the last stage is realized and there is a strict linear order between them. This situation is partly due to the fact that this theory was not fully taken advantage of by these studies since no specific references on it were included in their studies, although some earlier researchers such as DeRemer (1998) and Freedman and Calfee (1983) used to talk about it.

In view of the above, it is necessary for us to initiate research on raters' cognitive processes that are specifically grounded in the specific rating context of TEM 8 from the perspective of problem-solving theory. More specifically, we are urgent to know how raters build text images and how they articulate scores for TEM8-integrated compositions. To this end, the categories or labels of the rating process should be extracted from the field data rather than directly transplanted from other research. In short, the two research questions for this study are.

> Q1: How do raters as problem solvers realize the sub-goal of building text images in the TEM8 rating context? Or what operators are used for this sub-goal?

Q2: How do raters as problem solvers realize the sub-goal of articulating scores for text images in the TEM 8 rating context? Or what operators are used for this sub-goal?

Based on the results of these two questions, a model that is more in line with problem-solving theory is expected to be finalized. To this end, the present study relies on the verbal protocol analysis (VPA) methodology (Green, 1998; Charmaz, 2014), which is a qualitative method in which persons are asked to 'think aloud' and the researchers infer the cognition from the verbalization. This method is also highly recommended in cognitive psychology to describe the operators for solving the problem (Frensch & Funke, 1995).

Methods

TEM8 samples and profile of rating scale

We collected the writing scripts from 139 fourth-year undergraduate English major students from five intact classes in two national key universities in China. The course teachers were contacted and their students were assigned a timed reading-to-write task of the TEM8 test (see Supplementary I). As mentioned in the "Introduction" section, the TEM8 writing task meets the requirements of an integrated task. The candidates were given two excerpts on perfectionism over 350 words in total, one entitled "Headmistress Tells Pupils Not to Fret about Exams" and the other "The Pursuit of Perfection." They were asked firstly to summarize the main arguments in the excerpts and then to express "your opinion on perfection, especially on whether aiming for perfecting matters in whatever you do." They were allowed to use information from the excerpts to support themselves again in writing the argument section.

As most writing rating scales (Weigle, 2002), the descriptors of the highest level for each dimension of the TEM 8 rating scale reflect the full-fledged features of writing performance, while the other levels of descriptors remain the same substance, but the modifiers and qualifiers gradually decrease in degree. For example, the first descriptor in the content dimension of level 10-9 is "can accurately express the theme of the excerpts" while the 8-7 level is "can express the theme of the excerpts," and the level 6-5 is "can roughly express the theme of the excerpts." However, the TEM8 rating scale has three distinctive features as mentioned in the "Introduction" section, which can be shown in Supplementary II. It is by using this specific scale that the six raters completed their rating work.

Raters

Convenience and purposive sampling methods (Miles et al., 2014) were used to recruit study participants. Six highly experienced raters (five females and one male) agreed to participate in the study. They were from the School of Foreign Languages at two major universities in China. As shown in Table 2, four of them held Ph.D. degrees and two owned Master's degrees. Their average years of teaching experience was 19 years (min = 18; max = 23; SD = 2) with similar academic expertise (English pedagogy).

Table 2 Information of participants for a think-aloud experiment on rating

No	Name[a]	Degree	Teaching experience	Academic expertise
1	Hui	PhD	18 years/Comprehensive English	Cognitive linguistics/English pedagogy
2	Rui	PhD	20 years/Practical Writing	English pedagogy
3	Damei	Master	18 years/Comprehensive English	English pedagogy
4	Migrate	Master	23 years/English Literature	English pedagogy/EFL writing assessment
5	Marttew	PhD	18 years/Comprehensive English	English pedagogy/EFL writing assessment
6	Lei	PhD	18 years/Comprehensive English	English pedagogy

[a] Pseudonyms used

Table 3 Specifications of data collected from 6 raters

	Hui	Rui	Migrate	Damei	Marthew	Lei	Total
Length of TAPs/minutes	116	92	82	85	81	119	575
Length of interview/ minutes	22	19	16	13	7	23	100
Number of English words of TAPs	3591	3587	4756	4765	4629	3888	25,207
Number of Chinese characters of TAPs	14,263	12,088	11,317	9120	18,185	19,654	84,627
Number of English words of review	127	92	51	106	85	134	595
Number of Chinese characters of review	3175	3136	4044	2353	2588	4054	19,350

Procedure of VPA for the present study

In VPA method, a distinction in procedures for data collecting is often made between concurrent "think aloud" and retrospective "think aloud" (Green, 1998). Given the research questions, the present study adopted the former mode. It is believed that thinking in working memory is just available for only a very short time after it is experienced; thus, concurrent rather than retrospective "think aloud" data were collected (Barkaoui, 2011; Lumley, 2005). Prior to the "think aloud" experiment, all raters were informed of the purpose of the study and had an average of 10-min one-on-one training (see Supplementary III), during which they were trained to keep themselves from explaining or interpreting their thinking during reporting (Barkaoui, 2011). During the session, each rater was required to score 10 same writing samples and they could freely use either English or Chinese for reporting. After a concurrent "think aloud" experiment, we then conducted the interview. The three questions were listed in Supplementary III and were designed to elicit retrospective verbal reports of how they used the different points on the rating scales, for providing supplementary data of concurrent "think aloud" data (Green, 1998). As Table 3 shows, the average duration of the concurrent verbal report was 95.8 min (min = 81; max = 119; SD = 17.2), and the average duration of the retrospective interview was 16.7 min (min = 7; max = 23; SD = 6.0). We translated all the oral materials into text, and the length of the whole text was 103,977 Chinese characters and 25,802 English words.

Rating quality based on quantitative data

After the "think aloud" experiment, the six raters were asked to independently rate the remaining 129 samples within 4 h. For the scores of all the 139 writing samples generated by the six raters, we used the Multi-Facets Rasch Measurement (MFRM, Linacre, 2005) to analyze the fitness of the data, the purpose of which was to assess the quality of their

rating work. The main indices were as follows: (1) The separation index for students' ability was 4.57, indicating that students' abilities could be divided into five levels and the chi-square test results showed that (chi-square = 158.6, $p < 0.001$) there was a significant difference in the ability of the examinees. These results proved that the raters had the ability to discriminate the competence of the candidates. (2) The Infit MSq for all six raters were all within the range of 0.5 to 1.5, demonstrating good intra-rater reliability (Linacre, 2005) for all six raters, which demonstrated that for each individual rater, they could apply the same standards during the whole rating process consistently. (3) Exact agreement reflects the inter-rater agreement coefficient between raters, which is on the consistency of them in scoring the same composition (Linacre, 2005). In this study, the full agreement of the six raters was 36.8%, which was slightly lower than expected from the model (37.5%), indicating that the six raters had good inter-rater agreement but were independent of each other (Linacre, 2005). In addition, according to the results of the Kendall's Coefficient of Harmony analysis, except for the language dimension (Kendall's $W = 0.193$), the raters showed moderate correlation in content (Kendall's $W = 0.317$), structure (Kendall's $W = 0.361$) and total score (Kendall's $W = 0.303$), which further indicated that the consistency among the raters was reasonable. It was proved that they were competent raters, being able to discriminate ability levels of candidates and to rate consistently.

Immediately after they completed the rating, they were asked to finish the Confidence Level Questionnaire for Articulating Scores (Supplementary IV). The necessity of this questionnaire was twofold. On the one hand, the rating process validation framework (Knoch & Chapelle, 2018) asserts that the more confidence raters have, the more valid the rating process. On the other hand, from a problem-solving perspective, it is also necessary to know the extent of confidence, as it can tell us how problem solvers assess their problem-solving effectiveness (Ormrod, 2012). The questionnaire required the six raters to respond from 1 (not confident at all) to 4 (very confident) to show their confidence level. The result showed that raters expressed the highest confidence when rating organization (mean = 3.86; min = 3; max = 4; SD = 0.32) and almost similar levels of confidence when rating language use (mean = 3.53; min = 2; max = 4; SD = 0.63) and content (mean = 3.30; min = 2; max = 4; SD = 0.65). In general, they were positive about their efforts with high confidence in their work. Overall, the above results of both the Multifaceted Rasch analysis and questionnaire proved that they were rather qualified raters. Therefore, the verbal protocol report produced in this working environment could be an authoritative reflection of the rater's cognitive process, generalizable to raters of similar situations.

Data analysis

All the transcriptions (103,977 Chinese characters and 25,802 English words) were qualitatively analyzed for emerging themes. The data analysis could be divided into two phases. The first phase was segmentation where the transcription was divided into different segments, each of which represents one single process as a "unit of meaning" (Green, 1998). In this study, the two researchers were also assisted by the pauses in the reporters' speech flow, which provided cues for segmentation boundaries. The two researchers first independently segmented the transcription of the first rater, and their

consistency coefficient reached 0.85, which met the requirements for qualitative analysis (Green, 1998). On this basis, the two researchers each did half of the rest of the segmentation work. In total, 943 segments were identified.

In the second phase, the two researchers repeatedly read these fragments with the aim of generalizing and extracting themes that reflect the scoring processes. To this end, the researchers used a combination of top-down and bottom-up methods. Top-down means that the researchers considered the rating processes as operators or strategies to realize the sub-goals of the rating activity as a problem: building text image and articulating scores for text image, by following the principles of problem-solving theory of cognitive psychology. Bottom-up means that in determining the names of specific strategies or processes, the researchers were not constrained by the process names in models listed in the literature review, but followed the principle of direct induction from field data. Although some names might be similar to the models in the literature review, their connotation and denotation would be different in the present study. The extracted names and definitions on their own are the research result of qualitative analysis since they themselves have theoretical value (Charmaz, 2014), which is the main concern for a qualitative research of language testing by using the VPA method (Green, 1998). At the same time, in order to describe more patterns of rating behavior, we added up the frequencies of segments indicating different processes.

Results

Q1: How do raters as problem solvers realize the sub-goal of building text images in the TEM8 rating context? Or what operators are used for this sub-goal?

Operators for sub-goal I: building text images as isolated nodes

Qualitative analysis revealed that the sub-goal of building text images was realized by the raters by decomposing it into two sub-goals: building text images as isolated nodes and building holistic text for each dimension. In order to realize the first sub-goal, raters resorted to three operators or strategies: (i) single foci evaluating, (ii) diagnosing, and (iii) comparing, as shown in Table 4. Single foci evaluation means raters attended to a series of specific features of text, in which various nodes of isolated text images were built, without being further processed. Diagnosing is defined as identifying the shortcomings of compositions by pointing out how it should have been written by invoking theories of experts or raters' own knowledge of writing instructions. Comparing is defined as the process in which the text image of one composition is contrasted with the image of another.

Table 4 Operators for building text image as isolated nodes

Operators/processes	Segments	Freq.
1. Single foci evaluating	*The summary of this article is too short compared with the discussion part.*	646
2. Diagnosing for building text image as an isolated node	*According to my views, the names of key figures in the excerpts should be cited, which is the conventional way for summary writing.*	16
3. Comparing for building text image as an isolated node	*This composition uses more link words than the last one.*	13

Table 5 Sub-processes of single foci evaluating

Content	Freq.	Organization	Freq.	Language use	Freq.
1. Length between the summary and the argumentation parts	30	8. Arrangement of paragraphs of the whole passage	22	14. Use of vocabulary in the summary part	17
2. Completeness of points in the summary part	47	9. Inner structure of the summary part	25	15. Use of grammar in the summary part	18
3. Clarity of content in the argumentation part	103	10. Link words in the summary part	9	16. Use of vocabulary in the argumentation part	58
4. Details in the argumentation part	37	11. Logical relationship of adjacent sentences in the argumentation part	37	17. Use of grammar in the argumentation part	57
5. Use of argumentative strategies	72	12. Logical relation beyond adjacent sentences in the argumentation part	50	18. Use of punctuation in the whole passage	5
6. Use of evidence in the argumentation part	12	13. Link words in the argumentation part	24		
7. Naturalness of conclusion	23				

From the 646 segments indicative of the process of single foci evaluating, we further extracted 18 specific categories or processes indicating the specific aspect of textual features attended to by raters in the TEM8 rating context, as shown in Table 5. From Table 5, we can see that these processes can be grouped into three broader categories, each corresponding to the dimensions of the TEM8 rating scale: content (processes 1–7), organization (processes 8–13), and language use (processes 14–18), respectively.

Operators for sub-goal II: building holistic text image for each dimension

After realizing the sub-goal of building text images as isolated nodes, raters entered into a new stage where they presented a new problem. According to the requirement of the rating scale, raters were expected to assign scores for three dimensions (i.e., content, organization, and language use) rather than for each text image as single nodes. To shorten the distance between the current situation and the goal state, raters set a new sub-goal for this new state: to build the holistic text image for each dimension. In order to realize this sub-goal, raters resorted to two operators: (i) synthesizing and (ii) comparing, the exemplar segments and frequencies of which can be shown in Table 6. Here, synthesizing is defined as the process of deciding one dominant opinion for each dimension by weighing and integrating the text image as isolated nodes built in the last stage. From the exemplifying segment in Table 6, we can infer that through the synthesizing process, the rater transcended the boundaries among specific nodes and formed one holistic affirmative evaluation of content dimension as a whole. In this state, the process of comparing was also adopted but it was used for building holistic text images for each dimension rather than for text images as isolated nodes.

Q2: How do raters as problem solvers realize the sub-goal of articulating scores for text images in the TEM rating context? Or what operators are used for this sub-goal?

After raters built the holistic text image for each dimension, they entered into a new stage. Faced with the lack of specification on the alignment between the score band and the specific kind of holistic text image, they had to think about creative ways to solve the

Table 6 Operators for building holistic text image for each dimension

Operators/processes	Segments	Freq.
1. Synthesizing	*The whole composition has a well-written thesis statement, but some concrete content is fragmented. A good topic sentence can be identified in some paragraphs. But it lacks a sufficient summary of original excerpts. On the whole, it can reach the middle level in content.*	160
2. Comparing for building holistic text image	*The level of language use in this composition is clearly superior to that of the last one, and they are in different levels.*	17

problem of articulating scores for the holistic image of each dimension. To fulfill this sub-goal, they mainly adopted the following two groups of operators, as strategies for articulating scores: (i) demarcating borders between scores within one dimension and (ii) discriminating among dimensions for scoring, the definitions, exemplar segments, and frequencies of which can be shown in Table 7. Altogether, 91 segments indicative of the two groups of strategies as a whole were identified. According to Table 7, demarcating borders between scores refers to the strategy by which they purposefully allocated scores of the whole range to the holistic text image constructed by themselves. It includes three specific strategies: setting the baseline for articulating scores, classifying samples into level groups, and avoiding extreme scores. Discriminating among dimensions was adopted where raters had difficulties with disentangling the relations among the three dimensions. It also includes three strategies: differentiating between dimensions; simplifying and balancing.

Up to now, the two main questions of the present study have been answered. By arranging the above sub-goals, operators for the sub-goals, and their relationship into one organic whole, a finalized diagram reflective of how raters cognitively solved the problem of TEM8 rating activity was drawn, as shown in Fig. 3.

The most striking feature of the new model as in Fig. 3 is the much clearer representation of the linear relationship between processes, which is more in line with the problem-solving theory. As shown in Fig. 3, the raters, as problem solvers, generally divide the task into three successive states or stages, each with clear sub-goals and operators. After realizing the first sub-goal (building text image as isolated nodes), the raters enter the next state by setting another new sub-goal (building holistic text image for each dimension) before entering the last state with another sub-goal (articulating scores for holistic text image). The first state paves the way for the second state, which in turn becomes the starting point for the third state. By realizing each sub-goal in each state, raters continually reduce the difference between the status quo and the final state (Anderson, 2015).

Discussion

Firstly, the research findings on Q1 deepen our understanding of the role of building text images. As mentioned above, the concept of text image could be traced back to Freedman and Calfee (1983), and in their model, the three stages were represented as (1) reading and comprehending text to create text image, (2) evaluating text image and storing impressions, and (3) articulating evaluation. Literally, what was emphasized were: reading, understanding, evaluating, and articulating, rather than the text image itself, which put the function of the text image in a backstage position. In Wolfe's (2005) model, the

Table 7 Operators for articulating scores for holistic text image for each dimension

Group	Processes	Definitions	Exemplar segments	Freq.
Strategies aimed within one dimension	1. Setting baselines	Assigning a particular and personal meaning to a score, used as a benchmark	*I set 4 as the standard representing pass level, the compositions better than 4 can be given 5 or 6.*	18
	2. Classifying samples into level groups	Placing the scores into a broader scale before considering the more granular TEM8 rating scale	*I firstly classified the compositions into three broad bands as 'high quality group', 'medium quality group', and 'low quality group', respectively, then decided on concrete score within each broad bands.*	12
	3. Avoiding extreme scores	Avoiding controversy and complaints by not scoring too high or too low	*I tended not to give too high score since it was risky, especially on Language use dimension, and it is impossible for EFL students to obtain full score in this dimension.*	8
Strategies aimed between dimensions	4. Differentiating between dimensions	Disentangling the unevenness of writing quality in different dimensions	*There is very big problem on the structure of this composition, but its language has no serious problem.*	29
	5. Simplifying	Reducing the burden of assessing a particular dimension by considering its relationship to other dimensions	*Since there are only 3 points for organization dimension, for this dimension, I just look at some macro-level feature. Some problems in micro-organization will be put into content organization for subtracting scores.*	13
	6. Balancing	Reallocating scores between dimensions, taking into account their distinctiveness and the discrepancy in scores	*Since I subtract too many scores on language use dimension for it, I would take off the factor of language when scoring the content dimension.*	11

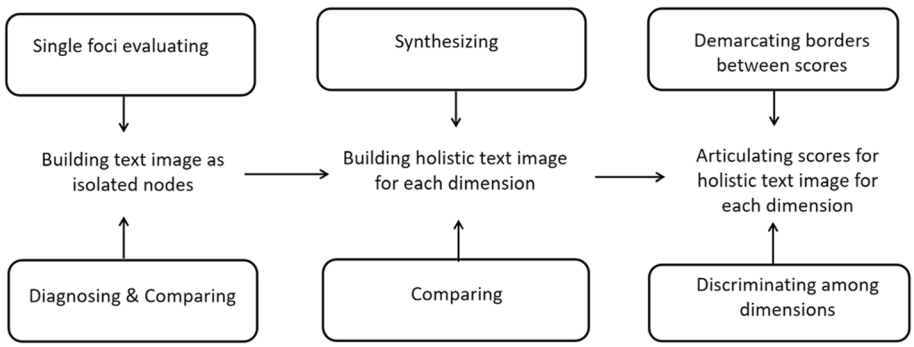

Fig. 3 Working model for the TEM8 rating process from the perspective of problem-solving

text image element was highlighted as an element connected to all other cognitive processes by arrows, but their interrelationship was not fully explained. In Lumley's (2005) model, the element of text image formation was even omitted, as analyzed above. Unlike them, the present study finds that the process of building text images should be viewed

more analytically, in which building text images as isolated nodes and building holistic text images should be viewed as two distinct sub-goals of two different stages. These two stages are chronologically distinct and the sequence between them is irreversible, as shown in Fig. 3.

This new treatment on building text images represents one originality of the present study, which is due to the introduction of problem-solving theory in cognitive psychology. Cumming et al. (2002) coding scheme placed all language-focused judgment strategies on the same level, as shown in Table 1. However, the last strategy named as *rate overall language* is not necessarily at the same level as the others, since it mainly serves to build a holistic text image, while the others serve to build text image as specific nodes, as suggested by the result of the present study. Similarly, the taxonomies on processes for building text images in Yan and Chuang (2023) also neglected the synthesizing process, a crucial link between text images and scores. In this sense, the present study provides a more complete picture of the mental process of rating. In addition, compared to the findings of Wolfe et al. (1998) that expert raters focused more of their attention on general text features while novice raters focus more of their attention on specific text features, we add that raters' cognitive behavior is more conditioned by the requirements of the type of rating scale. As shown in Fig. 3, even more competent raters as in this study have to experience the states of attending to specific textual features by single-focus rating, without which, it is impossible to move to the next state. For TEM8 composition raters, the mental process of attending to specific features of compositions is mandatory rather than optional.

Secondly, the research findings on Q2 deepen our understanding of the strategies used to articulate ratings. Compared with the classification of strategies on articulating scores (Lumley, 2005; Zhang, 2016), it is the present study that firstly makes a distinction between two main groups of articulating strategies: strategies aimed within a rating dimension and strategies aimed between dimensions, as shown in Table 7, which is reasonable and brings convenience for future research. With regard to the second group of strategies, Cumming (1990) used to mention that it posed a great challenge for raters to disentangle the different rating dimensions, and Marsh and Ireland (1987) even doubted that raters were actually able to discriminate between textual features of different dimensions, as a counter-argument to the use of analytical rating scales. From the perspective of rating cognition, the present study provides additional evidence that raters do indeed have the ability to discriminate between writing quality of different dimensions, as shown in Table 7. In addition, the processes defined as simplifying and balancing in the present study, which were ignored in the previous study, are a more sophisticated reflection of the raters' ability to discriminate between rating dimensions, which is partly caused by the unevenness of the scores for each dimension as a feature of the TEM8 rating scale. Raters are very agile in dealing with this particular rating context.

Thirdly, one of the merits of the newly constructed model is setting the relationship between processes in a more linear stage as shown in Fig. 3, which was not fully expressed in the previous models. As mentioned in the literature review, readers are unclear about the sequence of the specific processes in Wolfe's model (2005). This limitation is largely remedied in the new model. In addition, although the process of comparing was included in Wolfe's model (2005), its role was not elaborated. On the

contrary, the present model elaborates the function of comparing from two aspects: as operators for the sub-goal of building text images as isolated nodes and building holistic text images for each dimension, respectively. Besides, compared to Lumley's model (2005), the present model is more generalizable for describing the rating processes using analytical rating scales, since the analytical rating scale used in TEM 8 which includes many descriptors in one rating dimension is more common than the one used in STEP in Australia, as explained in the Introduction.

Fourthly, the present study provides evidence for the scoring validity of the TEM 8 integrated writing test in terms of the alignment between the raters' cognitive processes and the writing construct. On the one hand, what the raters paid attention to in state I (constructing the text image as isolated nodes) is quite consistent with theories of writing ability. For the six processes in Table 5 (numbered as 2, 3, 5, 12, 16, and 17), each of their frequencies exceeds 40, accounting for about 60% of the total frequencies, which represent the bulk of the cognitive activity used to evaluate the textual features. Their names as shown in Table 5 reflect a high degree of alignment with integrated writing skills as specified in the discourse synthesis theory (Spivey & King, 1989; Plakans, 2008; Gebril & Plakans, 2013), which include selecting, organizing, connecting, and expressing thoughts with appropriate vocabulary and grammatical forms. This alignment between them supports the scoring validity of the task (Knoch & Chapelle, 2018). On the other hand, the proportional relationship between the categories shown in Table 5 is consistent with the characteristics of presumed writing ability intended by the scale developers. According to the result of Table 5, if we add the frequencies within each broad category, we find that the total frequencies of processes in content, organization, and language use are 324, 167, and 155, respectively, with a corresponding proportion of 50.2%, 25.9%, and 23.9%. This result indicates that raters invested half of their cognitive effort in constructing textual images as isolated nodes in aspects of content. This result is in line with the intention of the developers of the rating scales: they considered content to be the most important dimension for judging writing ability by allocating half of the total score (10/20) to this dimension. In a nutshell, research into raters' cognitive processes provides valuable evidence of the scoring validity that cannot be obtained by solely relying on quantitative analysis of scores.

Implications

Firstly, based on the present study, we believe that it is high time to make a distinction between the two types of analytical rating scales: the analytical scale represented by TEM8 and the analytical scale represented by STEP (Lumley, 2005) and STAP (Mendoza & Knoch, 2018). In writing test research, the traditional distinction between holistic and analytical rating scales has been well established (Weigle, 2002; Fulcher, 2010; Bouwer et al., 2023), to the extent that the analytical rating scale was regarded as internally monolithic. On the contrary, the present study reveals that a paramount distinction should also be made between two types of analytical rating scales. In the former type, the rating criteria are quite broad, a criterion such as content including more than 10 descriptors, whereas in the latter type, a criterion usually consists of one descriptor. Because of these differences, these two types of scales are contrasted in terms of the scores meaning and functions of the scores. For the TEM8 analytical scale, the scores articulated

by the raters for each dimension should be seen as indicating the holistic ability of the candidates for each dimension as a whole, because it is the holistic text image formed by the synthesizing process rather than the image of isolated nodes that are assigned scores by the raters, as shown in the present study. Functionally, therefore, such dimensional scores are not suitable for diagnostic purposes. For example, if a candidate receives a score of 5 in the language use criterion, it just represents the level of general ability of using language, but it does not correspond to any specific aspects such as vocabulary as a single construct or grammar as a single construct. On the contrary, in analytical scales such as STEP and STAP, the meaning of the dimensional score can be seen as representing a more specific textual feature, and thus, the scores have a more diagnostic function. In practice, analytical scales of the former type are more suitable for large-scale language proficiency testing, while the latter type is more suitable for classroom-based formative assessment. In short, this distinction has great value for writing assessing practices. In determining the meanings and functions of scores, test practitioners should have in mind the types of analytical rating scales.

Secondly, the present study allows us to explain rating competence by combining qualitative and quantitative results. On the one hand, we can claim that for each single rater, they have the ability to construct the text image and adopt the strategies for articulating scores, by which they can discriminate the competence of the candidates in a self-consistent way. This is the underlying reason for the quantitative data analysis result of indices such as separation ratio, separation index, and Infit MSq (see rating quality based on quantitative data in Methods section). On the other hand, the holistic text images constructed by different raters for the same composition are not necessarily the same (Wolfe, 2005), and furthermore, the strategies for articulating scores are unique for each rater, which makes them produce related but different scores for the same composition. This is the underlying reason for the fact that there is a moderate Kendall's coefficient of harmony shown in the "Methods" section. On this basis, the empirically constructed working model shown in Fig. 3 can provide a reference for taxonomies describing the components of rating competence, skills that can be acquired by novice raters through training (Yan & Chuang, 2023). For example, we can list the competence of diagnosing, the competence of synthesizing, the competence of demarcating borders between scores, etc., for a more complete description of rating competence.

Conclusion

In general, this paper fills the gap that we lacked knowledge about how raters cognitively solve the problem of scoring compositions of TEM8. For research Q1, we have found that the goal of solving the problem of building text images was decomposed into two sub-goals: building text images as isolated nodes and building holistic text images for each dimension. To realize these sub-goals, raters mainly resorted to the following processes: single foci evaluating, diagnosing, comparing, and synthesizing. For research Q2, we have found that the goal of articulating scores for a holistic text image was realized by two operators or groups of strategies: demarcating boundaries between scores within a dimension and discriminating between dimensions for scoring, each involving more specific strategies. Based on these findings, a working model symbolizing the rating process of TEM8 integrated writing was finalized.

To the best of our knowledge, the model constructed in Fig. 3 is the first to be designed for the scoring context of an integrated writing test with an analytical scale such as the TEM 8 style. Through these findings, we have enriched our knowledge in this field in several aspects. For example, the rules and functions of building text images are elaborated more systematically than in previous research; a more systematic classification of score articulation strategies is extracted in the present study. In addition, the validity of the scores was supported by the results of qualitative analysis. The present study implies that a clear distinction should be made between two types of analytical rating scales, which have long been overshadowed by the distinction between holistic and analytical scales. Besides, the implications for rating competency were provided.

This study is not without its limitations. One limitation is the uneven granularity of labels for analyzing the verbal protocol data, in that for the single focus rating segments, we further divided them into 18 categories, but synthesizing as a process is rather broad without further decomposition, which can be resolved by future research. More raters from heterogeneous groups are needed if future comparative studies of rating processes are to be undertaken. In addition, a large number of raters will be needed if quantitative analyses are to be carried out in the future to test this newly constructed model, for example, to test the validity of the model by using the method of confirmatory factor analysis.

Abbreviations

CET 4	College English Test band 4, China
CET 6	College English Test band 6, China
EFL	English as a foreign language
ESL	English as a second language
IELTS	International English Language Testing System
STEP	Special Test of English Proficiency, Australia
TEM8	Test for English Majors Band 8, China

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-023-00265-x.

Additional file 1: Supplementary I. The TEM8 integrative writing task. **Supplementary II.** Rating scale for TEM8 integrative writing test. **Supplementary III.** Instructions of concurrent TAP & interview questions. **Supplementary IV.** Confidence level questionnaire for articulating scores.

Acknowledgements

The authors wish to express their gratitude to all 6 raters who kindly helped with data collection in this study.

Authors' contributions

Dr. Jia was involved in the data collection, analysis, and writing of the early drafts. Dr. Zhang also participated in the data analysis, and her main job was to revise the earlier drafts. The authors read and approved the final manuscript.

Authors' information

Dr. Wenfeng Jia is a lecturer at the College of Translation Studies at Shandong University, China. His research interests include writing assessment and English pedagogy. His recent publications have appeared in peer-reviewed journals such as *Modern Foreign Languages*, *Foreign Language Testing and Teaching*, and *Journal of Sichuan International Studies University*.

Dr. Peixin Zhang is currently an associate professor in the College of Foreign Languages and Cultures at Xiamen University, China. Her research interests are in the areas of translation assessment, writing assessment, and research methods. Her recent publications have appeared in peer-reviewed journals such as *Assessing Writing*, *Target*, *Foreign Language Teaching and Research*, and *Modern Foreign Languages*. She is the author of *A Study of Analytic Rating Scale for Chinese-to-English Translation Competence Test in Translation Teaching* (Xiamen University Press, 2017). ORCID: https://orcid.org/0000-0002-2258-0451

Funding
This work was supported by Fujian Province Philosophy and Social Science Planning Project (FJ2021B113).

Availability of data and materials
The data associated with this study would be available upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 5 July 2023 Accepted: 17 October 2023
Published online: 26 October 2023

References

Anderson, J. (2015). *Cognitive psychology and its implications* (8th ed.). Freeman.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001
Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418
Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75. https://doi.org/10.1177/0265532210376379
Bouwer, R., Koster, M., & Van den Bergh, H. (2023). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy & Practice*, *30*(3–4), 302–319. https://doi.org/10.1080/0969594X.2023.2241656
Charmaz, K. (2014). *Constructing grounded theory*. SAGE.
Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700103
Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 216–229). Wiley.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137
DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*(1), 7–29. https://doi.org/10.1016/S1075-2935(99)80003-8
Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. http://www.jstor.org/stable/356600
Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). Longman.
Frensch, P. A., & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 25–44). Lawrence Erlbaum Associates.
Fulcher, G. (2010). *Practical language testing*. Routledge.
Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, *10*(1), 9–27. https://doi.org/10.1080/15434303.2011.642040
Green, A. (1998). *Verbal protocol analysis in language testing research: a handbook*. CUP.
Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia*, *12*(20), 1–19. https://doi.org/10.1186/s40468-022-00168-3
Jacobs, H. L., Zingkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Newbury House Publishers, Inc.
Kluwe, R. H. (1995). Single case studies and models of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 275–295). Lawrence Erlbaum Associates.
Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 1–23. https://doi.org/10.1177/0265532217710049
Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing*, *18*(4), 300–308. https://doi.org/10.1016/j.asw.2013.09.003
Li, H., & He, L. Z. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, *12*(2), 178–212. https://doi.org/10.1080/15434303.2015.1011738
Linacre, M. (2005). *A user's guide to facets*. MESA Press.
Liu, B. Q., & Fan, J. S. (2020). Teachers' views on the high stakes language assessment reforms: The case of Test for English Majors Band 8 (TEM8). *Foreign Language World*, *43*(2), 20–26.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa
Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.
Marsh, H. W., & Ireland, R. (1987). The assessment of writing effectiveness: A multidimensional perspective. *Australia Journal of Psychology*, *39*, 353–367. https://doi.org/10.1080/00049538708259059
McNamara, T. (1996). *Measuring second language performance*. Blackwell.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, *35*, 41–55. https://doi.org/10.1016/j.asw.2017.12.003
Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: a methods sourcebook* (4th ed.). SAGE.
Ormrod, J. E. (2012). *Human learning* (6th ed.). Prentice-Hall Inc.
Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111–129. https://doi.org/10.1016/j.asw.2008.07.001
Spivey, N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly*, *24*(1), 7–26. https://www.jstor.org/stable/748008
Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*(1), 37–56. https://escholarship.org/uc/item/83b618ww
Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002
Yan, X., & Chuang, P. L. (2023). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 153–179. https://doi.org/10.1177/02655322221074913
Zhang, J. (2016). Same text, different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37–53. https://doi.org/10.1016/j.asw.2015.11.001

Publisher's Note