ORIGINAL RESEARCH
published: 26 April 2018
doi: 10.3389/feduc.2018.00027

Measuring L2 Essay Rating Strategy: Scale Development and Preliminary Validation

*Ying Xu**

Department of College English, School of Foreign Languages, South China University of Technology, Guangdong, China

Given the lack of an L2 essay rating strategy scale, this study aims to develop and evaluate an instrument measuring College English Test Band 4 (CET4) essay raters' rating strategy. 14 raters were first invited to mark 10 mock essays while conducting think-aloud in order to generate a pool of scale items. After piloted on three raters, the initial version of the questionnaire including 28 items was established, among which 22 items were related with cognitive strategies (CS) and the rest meta-cognitive strategies (MCS). Then it was administered to a sample of 450 raters in four marking centers around China. Item-total correlation, exploratory factor analysis (EFA), and reliability analysis were conducted to evaluate its psychometric properties. It was found that the final version contained 12 items, and the four-factor solution (*Self-evaluate*, *Decide*, *Diagnose*, and *Compare*) could explain 59.92% of the total variance. Cronbach's Alpha reached 0.73 for the whole questionnaire. The results suggest that the instrument has acceptable reliability and adequate validity.

Keywords: CET4, rating strategy, cognitive strategy, meta-cognitive strategy, think-aloud, questionnaire

Edited by:
Ernesto Panadero,
Universidad Autonoma de Madrid,
Spain

Reviewed by:
Carmen Tomas,
University of Nottingham,
United Kingdom
Gustaf Bernhard Uno Skar,
Norwegian University of Science and
Technology, Norway

Correspondence:
Ying Xu
xuying@scut.edu.cn

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *11 December 2017*
Accepted: *12 April 2018*
Published: *26 April 2018*

Citation:
Xu Y (2018) Measuring L2 Essay
Rating Strategy: Scale Development
and Preliminary Validation.
Front. Educ. 3:27.
doi: 10.3389/feduc.2018.00027

INTRODUCTION

The field of language testing has switched its attention to performance assessment since the communicative turn in the early 1980s (McNamara, 1996). As the key factor in performance assessment, raters are supposed to provide appropriate ratings and do so in a consistent way (Lim, 2011). The implication is two-fold: on the one hand, raters should provide reliable scores which are accurate, reproducible, and generalizable to other testing occasions and other similar test instruments (Ebel and Frisbie, 1991); on the other hand, raters should carry out scoring in a manner consistent with the construct and measurement goals in order to support a validity argument for scores (Bejar, 2012). Needless to say, rating test-takers' performances has occupied a central position in language testing because it not only carries significant practical implications, but also determines the validity of test interpretation and use (Crisp, 2012; Knoch and Chapelle, 2017).

Because rating the performance of a test-taker is essentially "a complex and error-prone cognitive process" (Cronbach, 1990, p. 584), raters may employ various cognitive strategies (CS), from retrieving information in the memory, selecting, weighing, to integrating information in order to generate a score, during which they are influenced by factors like their prior experience and personal backgrounds (Baker, 2012). As a result, the perennial problem of rater variability becomes a major threat to the validity of the inferences drawn from the test results. Rater variability was found in diverse forms, like the degree of raters' scoring severity or leniency, the degree of consistency of their ratings across different facets (e.g., test-takers, scoring criteria, and tasks), the degree of their compliance with the rubric, the way that they interpret criteria, and the strategy taken to reach a final score (Bachman and Palmer, 1996; McNamara, 1996; Weir, 2005).

Traditional research on rater variability focuses on different demographic variables such as raters' L1 background (Kobayashi, 1992; Connor-Linton, 1995; Kobayashi and Rinnert, 1996; Shi, 2001; Johnson and Lim, 2009; Kim, 2009), their academic and educational backgrounds (Brown, 1991; Cumming et al., 2002), and their teaching and rating experience (Cumming, 1990; Shohamy et al., 1992; DeRemer, 1998). Among these factors, rating experience seems to be a major one which could determine rater expertise in the ESL essay rating process (Suto, 2012) because the concrete experience is necessary to help raters build a well-organized knowledge structure and ultimately develop domain-specific expertise. Therefore, it is heavily investigated in the L2 essay rating literature. For example, Lumley (2005) adopted six variables to identify highly trained and experienced raters: experience with the STEP (Special Test of English Proficiency), similar STEP training and experience, experience with other tests and assessment procedures, educational background, teaching experience, and internal consistency measured by infit mean square value.

The last three decades have witnessed a shift of research focus to rater cognition, which is conceptualized as being "… concerned with the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (Bejar, 2012, p. 2). An increasing amount of research purports to investigate how raters make a judgmental decision at a particular rating because as Connor-Linton (1995) pointed out, "… if we do not know what raters are doing, then we do not know what their ratings mean" (p. 763). The other reason for the shift is that individual differences in raters' cognitive behaviors could possibly account for unexplained, systematic rater variability that resists training (Wolfe et al., 1998; Weir, 2005; Baker, 2012). Therefore, various models of rater cognition have been established. Methodologically speaking, most research so far follow the qualitative research paradigm by adopting methods such as think-aloud protocols (TAPs, see a comprehensive review by Barkaoui, 2011), interviews (e.g., Milanovic et al., 1996), stimulated recall (e.g., May, 2009), and written comments (e.g., Shi, 2001; Yan, 2014), yet quantitative methods like questionnaire were seldom used.

However, as most qualitative studies on rater cognition employ a small sample of raters, the generalizability of research findings is challengeable, as admitted by Vaughan (1991). No doubt that compelling evidence of raters' cognition from a larger sample are badly needed in order to triangulate findings thus far. Therefore, the present study aims to fill the research gap by constructing a psychometrically sound instrument for assessing essay rating strategy in the CET4 context. CET4 is the largest and most influential language test in China (Jin, 2011), with more than 18 million test-takers annually (Zhang, 2016).

LITERATURE REVIEW

A review of research on rater cognition in direct writing assessment can identify two main foci: (1) the performance features which raters attend to and weigh when assigning ratings; (2) raters' CS and rating styles which may affect their judgments. The formal is a crucial issue related with score validity because whether raters attend to rubric-irrelevant features like handwriting and length of the essay (Vaughan, 1991; Lumley, 2005; Barkaoui, 2010), or whether they overweigh some criteria while undervalue others (Cumming, 1990; Cai, 2015) may cause threats to test validity. Besides, the latter is also a validity issue because it is concerned with the nature of rating expertise, which carries practical implications regarding rater training. As the present study is to measure CET4 essay raters' rating strategy, some typical models of rater cognition in the literature would be reviewed below.

The earliest model of essay scoring is proposed by Freedman and Calfee (1983), which suggests a linear rating process containing four stages: (1) reading text to build a text image; (2) evaluating the text image; (3) articulating the evaluation; (4) monitoring the process. The text image, conceptualized as the mental representation of an essay that is created as raters read and interpret the essay, is at the core of this model. Since raters interpret the essay on the basis of their own background knowledge, beliefs and values, and knowledge of the writing process, the text image may be different from one rater to another. Moreover, it may not be an exact replication of the original essay. Based on the text image, raters are able to compare different features of writing with their mental representations of the scoring criteria. In this process, raters make judgments on the quality of the text, and formulate decisions. Drawing on insights within cognitive psychology, this model manages to theorize rater cognition by "… conceptualizing human raters as 'information processors' undertaking 'multiple attribute' judgments" (Suto, 2012, p. 22), and has exerted a profound influence in the field.

Vaughan (1991) used TAPs to investigate nine raters' thought process, and found that raters generally followed the rating criteria in the rubric. But if there were any features in the script which did not fit pre-defined categories, raters would have to make decisions that were not based on the rubric or any training they received. In this case, raters would rely on their first impression or use one or two categories, like grammar and/or content, to give a score, showing a reductive nature (Rezaei and Lovorn, 2010). She finally argued that "Despite their similar training, different raters focus on different essay elements and perhaps have individual approaches to reading essays" (p. 120).

Milanovic et al. (1996) employed the retrospective written report, the introspective verbal report, and the group interview to triangulate data collection in a study of the decision-making behavior of 16 composition markers. They devised a model of the decision-making involved in holistic scoring: raters first scan the script and form an overall idea of the length, format, handwriting, and organization; then they read quickly to establish an indication of the overall level of the writing script; finally they proceed to rating. Four reading styles were revealed: "the principled two-scan/read," "the pragmatic two-scan/read," "the read through," and "the provisional mark."

DeRemer (1998) viewed the rating process as a problem-solving process, and rating as a constructive activity. By analyzing three raters' verbal reports, she distinguished three kinds of rating cognitive operations: goal setting, evaluation, and relations. She found that raters did not simply make a match between their response to the text and descriptors in a rubric. Instead, they had to reconcile the interpretation of the rubric with the specifics of the text. As essay rating is recognized as an "ill-structured task" (DeRemer, 1998, p. 14), raters have to develop their own strategies to solve the problem.

Unlike the above studies which treated raters as a homogeneous group, Cumming's study (Cumming, 1990) compared decision-making strategies between seven novice raters and six expert raters. It was found that as a whole, raters used a wide range of knowledge and strategies, and their decision-making processes were complex and interactive. 28 interpretation and judgment strategies were identified and can be grouped into four kinds: (1) the focus on the raters' self-control of their own reading or judgment process; (2) the focus on the substantive content of the texts; (3) the focus on the use of language in the texts; (4) the focus on the rhetorical organization of the texts. These strategies varied significantly in use between the two groups. Expert raters developed a more comprehensive mental representation of the problem of essay evaluation, and used a large number of diverse criteria, self-control strategies, and knowledge sources to read and judge students' texts than novice raters. On its basis, Cumming et al. (2002) developed and validated a descriptive framework of the decision-making processes of raters as part of the development of TOEFL 2000. First, a preliminary descriptive framework was built up based on TAPs of 10 experienced EFL/ESL raters while rating without scoring criteria. Then, it was applied to verbal data from another group of seven experienced English-mother-tongue (EMT) raters. Last, it was revised by analyzing TAPs from seven of the same ESL/EFL raters. It was found that both rater groups employed a prototypical sequence of decision-making behaviors (i.e., first to scan for surface-level identification; then to engage in interpretation strategies; finally to articulate a scoring decision), and in similar proportions of frequency. There are two breakthroughs in this study. First, it established a useful taxonomy of raters' complex cognitive behaviors by introducing two parameters: focus and strategy. In this way, discussion of rater cognition can be held in terms of three foci (self-monitoring, rhetorical and ideational, and language) and two strategies (interpretation and judgment). Second, raters' meta-cognitive strategies (MCS) began to be separated from CS, and stand alone as an independent category in the literature.

Another line of research (Wolfe, 1997, 2006; Wolfe et al., 1998) extended Freedman and Calfee's (1983) seminal work, and developed a cognitive model of rater cognition (Wolfe, 2006) which incorporates a framework of scoring and a framework of writing. In this model, text image takes up a central position. The framework of scoring is a mental representation of the process through which a mental image of the text is created and the quality of that mental image is evaluated. At the stage of reading, raters may comment in a non-evaluative manner about their reactions to the context. At the stage of evaluating the text image, they may monitor specific characteristics of the text to determine the degree to which the text image demonstrates rubric. After reading, they may review the most noticeable features and decide scores. At the stage of justification, they may provide rationales for the scores, or diagnose how the text could be improved, or compare the essay with other texts. The framework of writing is a mental representation of the scoring rubric. The components of this framework may differ significantly from rater to rater, thus are developed based on raters' individual experiences as well as effort to train the rater to adopt the rubric.

Lumley (2002, 2005) used TAPs to investigate four raters' thinking process while using an analytic rubric. It was found that his raters would use their own knowledge or intuition to resolve eventualities not covered by rules and the scale, or take some strategies like attaching heaviest weight on one aspect or making comparison with previously rated essays. Different from the cognitive psychological tradition of research on rater cognition, Lumley (2005, p. 291) raised a socio-cognitive model of rating process, which included three levels (institutional level, instrumental level, and interpretation level) at which the process operated in three stages (reading, scoring, and conclusion). The highlight of this model is the idea of analyzing essay scoring at the institutional level, which should be seen as involving major socially motivated components. First, at the reading stage, administration of a test targets at eliciting the performance from test takers. Then, at the scoring stage, a number of institutional constraints come into play, such as the rubric, training, reorientation, and professionalism. Finally, at the conclusion stage, the institutional goal lies at the end of the operation of these institution constraints.

The latest model of rater cognition was proposed by Bejar (2012), which was built on the previous research on the scoring of constructed responses. A contribution of this model is to approach rater cognition using an argument-based approach to validation (Kane, 2006). It made a distinction between the assessment design phase and the rating process. Although the scoring phase bears great resemblance to Wolfe's (2006) model, Bejar (2012) regarded the process of rating as a loop, and argued that raters could develop a more productive strategy in order to "… minimize cognitive effort but gets the job done" (p. 6).

Another relevant study is Baker (2012), which examined rater decision making style (DMS) in a high-stakes writing assessment by using the mixed-method exploratory case study. It managed to create the DMS profiles of six raters by combining four data sources including write-aloud comments, Facets fit statistics, doubled scores and the general decision-making style inventory (GDMSI). The author concluded that individual sociocognitive differences in DMS could account for some rater variability in scoring.

Last, Zhang's (2016) study on CET4 essay raters' CS is also worth reviewing because it aims to explore how raters' use of cognitive and MCS influence rating accuracy in the context of CET4 essay rating in China. 13 CET4 essay raters were categorized into ACCURATE and LESS ACCURATE groups, and their TAPs were collected and analyzed drawing on the framework of Cumming et al. (2002). It was found that the ACCURATE group performed better at integrating information from target essays and had a better consciousness of their own rating accuracy. As the research context is also CET4 essay rating, the analysis of raters' verbal protocols provides a useful reference for the present study.

All of the above research provided a knowledge base of rater cognition research. Nevertheless, a close look at them could turn out two common features. First, all studies adopted a qualitative approach and employed a small number of raters (around ten), hence the generalizability of research findings was threatened. Second, the method of think-aloud protocols was used to collect data, although this technique is commonly recognized with two validity issues, veridicality (to what extent it can accurately represent the participants' true and complete thinking process) and reactivity (to what extent the report of the rating process can affect the process being observed and/or its outcomes).

In order to investigate into CET4 essay raters' CS and MCS, the present study intends to develop a CET4 Essay Rating Strategy Questionnaire (CERSQ) and test its reliability and validity. Cognitive strategy refers to human information processing that would "… involve mental manipulations or transformations of materials or tasks that serve to enhance comprehension, acquisition, or retention" (O'Malley and Chamot, 1990, p. 229). In the present study, it is operationalized as the raters' mental behaviors that are triggered in their decision-making process. In comparison, meta-cognition refers to "… the ability to think about one's own cognition and regulate it" (Suto, 2012, p. 23).

METHOD

Design

The study adopted a two-phase mixed methods sequential exploratory design (Creswell and Plano Clark, 2007). In Phase 1, qualitative data (14 CET4 essay raters' verbal protocols while rating 10 CET4 mock essays) were collected in order to construct questionnaire items; in Phase 2, quantitative data (450 CET4 essay raters' responses to the questionnaire) were obtained in order to validate the instrument.

Context of the Study

In China, essay rating in the large-scale national language tests is always a thorny issue because of the large number of test-takers and the limited rating time period. CET4 essay rating is no exception. According to the test specifications, CET4 is designed as a criteria-related and norm-referenced standardized test targeted at tertiary level non-English majors across the nation (Yang and Weir, 1998). It is criteria-related in the sense that the development of the whole test is based upon standards specified in the *College English Curriculum Requirements* (Ministry of Education, 2006). At present, CET4 and its sister (CET6) are the largest and most influential English tests in China. CET4 has two administrations each year, and each administration has nine million test-takers (Jin, 2017). The huge test population brings heavy workload to raters, hence "… it takes 2 weeks for over 4,000 raters in 12 marking centers across the country to complete the scoring of 9 million writing scripts and 9 million translation scripts after each test" (Jin, 2017, p. 12).

CET4 has a writing task to assess test-takers' writing ability, which accounts for 15% of the total score. Test-takers are supposed to write within 30 minutes a short composition of at least 120 words on a general topic or an outline. The essay should be basically complete in content, appropriate in diction, and coherent in discourse. An online marking system has been used since 2006, which can provide real-time statistics of raters' performances such as the mean score, the standard deviation, and the graphical score distribution. The most important criterion for quality control is the correlation coefficient between the essay scores raters have awarded and the total scores of the objectively rated items for the same examinee groups (Zhang, 2009). The rationale behind this criterion is a hypothesized linear relationship between students' writing skill and other language skills (represented by the total scores of the objectively rated items including listening comprehension and reading comprehension items). It can generally work well but is still facing some challenges because an increasing number of test-takers tend to memorize rote-patterns prepared by some test-preparation and coaching institutes, which could cause a mismatch between students' general language ability and their writing competence. As a result, raters sometimes are left puzzled in their judgments and the correlation coefficient would be distorted. Supervisors and directors in a marking center have access to the statistics. If they spot any aberrant raters, whether with a fairly low correlation coefficient or an abnormal low/high mean or standard deviation, they will decide an intervention to apply, ranging from reviewing the scoring guide, remarking the last 10 essays, to retraining.

Participants

Seventeen raters from the CET Guangzhou marking center voluntarily took part in the first phase of the study. 14 of them were trained to rate 10 essays with TAPs owing to their interest in the study, ability to produce verbal protocols during the rating process, and rich experience of rating CET4 essays. Furthermore, the rest three raters took part in the pilot study. **Table 1** shows the profile of 14 raters. As there are two CET4 essay rating sessions in a year, raters' experience was operationalized as the times of rating (TOR) of CET, namely the times that they have attended the CET4 essay rating session. According to **Table 1**, these raters' TOR ranged from 6 to 15 ($M = 10.57$, $SD = 2.53$), showing that these raters had rich experience in rating CET4 essays.

After piloted on the other three raters, the questionnaire was administered to a set of 450 CET4 essay raters from four marking centers in China (including Chengdu, Chongqing, Guangzhou, and Xi'an). After removing ambiguous data, missing data, and outliers in the data screening, the final sample comprised 367 teacher raters (75 men and 292 women) who taught English at universities. Their ages ranged from 27 to 50 years ($M = 38.26$, $SD = 6.74$), and their TOR was between 1 and 25 ($M = 6.14$, $SD = 5.36$). Among them, 113 raters partook the CET4 essay rating for more than 6 times ($M = 12.28$, $SD = 4.73$), which could be labeled experienced raters.

TABLE 1 | The profile of 14 raters who rated 10 essays with TAPs.

Rater	Gender	Age	Times of rating (TOR)
R1	M	32	6
R2	M	31	10
R3	M	37	12
R4	M	40	12
R5	F	36	10
R6	F	31	7
R7	M	35	12
R8	F	36	11
R9	M	34	13
R10	F	40	15
R11	M	39	12
R12	F	34	8
R13	M	31	12
R14	F	34	8

Materials and Instruments

As a result of unavailability of the test data, mock essays had to be used for the study. These essays were written by students at the author's university who had just finished the December 2012 CET4 administration. Students from six classes were asked by their English teachers to rewrite on the writing task (Appendix 1). A total set of 300 essays was collected. A sample mock essay can be found in Appendix 2 in Supplementary Material. After being marked by two CET4 essay rating experts, 10 essays with the exactly agreed-upon score covering the full score range were selected. Then they were used in January 2013 CET4 essay rating session as the prompt to elicit 14 raters' verbal protocols. Since the focus of the study is the rating strategy, using these essays can satisfy the research purpose. The 10 essays' scores are 9, 11, 2, 10, 8, 3, 7, 6, 14, 11 respectively.

CET4 writing adopts a holistic rubric (Appendix 3 in Supplementary Material), which is made up of five score bands. It describes four constructs including coherence, topic relevance, comprehensibility, and accuracy. It is an augmented rubric using five scores (2-, 5-, 8-, 11-, and 14-point) to anchor raters' mental representation of each score band. Therefore, the difference between varying score bands is thought to be substantial, e.g., 2-point and 5-point. In contrast, the difference within certain score band is deemed acceptable because it may be attributed to raters' random errors in judgment, e.g., 4-point and 6-point. In practice, raters are trained to first categorize each essay into one of the five score bands and then decide the final score by either adding or subtracting one point if the essay is perceived slightly better or worse than the range finders (i.e., five benchmark essays provided by National College English Testing Committee to anchor raters' judgment) in that score band.

TAPs was used to collect the introspective data, although the problems of veridicality and reactivity are identified (Barkaoui, 2011). To guarantee the validity of this method, several measures were taken: (1) a careful selection of raters; (2) training informants not to analyze their thought by demonstrating the think-aloud process with two mathematics problems (Ericsson and Simon, 1984); (3) providing raters with printed instructions that were clear and unambiguous following Ericsson and Simon (1984, p. 375) on how to conduct thinking aloud while performing the rating task; (4) prompting raters consistently during their think-aloud process if there were pauses longer than 10 seconds; (5) taking observation notes during the session if any raters kept indicating difficulty in following the procedure, then conducted a member check to see whether they perceived the method to be valid in eliciting their thought process. All verbal protocols were recorded with a digital voice recorder.

This study was reviewed and approved by the ethics committee affiliated to the School of Foreign Languages, South China University of Technology. The study was carried out in accordance with the recommendations of the ethics committee that approved the study. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Procedures

The scale items were generated based on the analysis of verbal protocols. First, 14 raters received a training of how to rate essays while conducting think-aloud. Then, they were provided with the prompt used in the December 2012 CET4 administration and five range-finders to get re-familiarized with the rating task. Finally, they were invited to rate the 10 essays with TAPs. Based on the coding scheme of TAPs, the initial version of CERSQ was drafted (Appendix 4 in Supplementary Material). It was piloted on three raters and no serious problem was identified. During the July 2015 CET4 essay rating session, it was administered to 450 raters. Data collection was embedded within the period of CET4 essay rating session on account of the availability of numerous raters.

Data Analysis

Regarding the qualitative data, 14 raters' verbalized thoughts were transcribed, segmented, and coded to determine their strategies in rating essays. First, the record of TAPs was transcribed verbatim by a research assistant, and all transcribed protocols were double-checked by the author. Second, the transcribed texts were segmented into independent "idea units" (Brown et al., 2005, p. 14). Finally, idea units were coded based on a priori theoretically motivated coding because studies reviewed above (e.g., Cumming et al., 2002; Lumley, 2005; Wolfe, 2006) have provided a useful reference. Most idea units can be coded directly based on previous findings. As for those newly emergent strategies, open pattern coding following the general concepts of Grounded Theory (Glaser and Strauss, 1967) was implemented. In total, the transcript of the verbal protocols was segmented into 1,210 idea units. As a reliability check, all data were coded by a CET4 essay rating expert (a PhD in Applied Linguistics) and the author separately. The two coders agreed on 1,120 among all units, indicating a satisfactory inter-coder agreement (92.56%). Disagreements were resolved through negotiation.

As for quantitative data, evidence for the reliability and validity of the instrument was collected by conducting item-total correlation analysis, exploratory factor analysis (EFA), and reliability analysis.

TABLE 2 | The results of exploratory factor analysis (EFA).

No	Item	Factors and their loads			
		Factor 1 (SEV)	Factor 2 (DEC)	Factor 3 (DIA)	Factor 4 (COM)
23	I reflect on my own leniency or harshness after I mark an essay.	0.838			
28	I introspect the sufficiency of the rationale for my score.	0.740			
14	I evaluate my confidence after giving a score.	0.706			
2	I predict the author's possible score after I read the beginning of the essay (i.e., the 1st paragraph).		0.789		
10	I don't decide the score until I finish reading the whole essay.		0.740		
18	I determine the score band first, and then the exact score.		0.658		
9	I check the gravity of errors during reading.			0.819	
21	I classify errors into different types for an essay.			0.727	
3	I diagnose the frequency of errors during reading.			0.671	
6	I associate the present essay with five range-finders.				0.745
25	I compare the essay with my own standard of essays with the passing score (i.e., the point of 9).				0.676
11	I make a comparison between the present essay with previously rated ones.				0.668
Eigenvalue		3.19	1.51	1.42	1.07
Percentage of variance		16.92	14.95	14.54	13.51

RESULTS

Results of Verbal Protocol Analysis

The coding scheme and frequency of each sub-category is shown in Appendix 5 in Supplementary Material. Two main categories (CS and MCS) were identified. The former consists of seven categories (*Interpret*, *Decide*, *Diagnose*, *Edit*, *Infer*, *Summarize*, and *Compare*), and the latter two categories (*Self-evaluate* and *Avoid*). Each category is made up of three sub-categories except *Interpret* (with four), so the total number of sub-categories is 28. For the main categories, the number of CS (941) surpasses that of MCS (269); for the categories, *Interpret* has the largest number of frequencies (336) while *Avoid* the least (60); for the sub-categories, *INT-2* is the largest (104) while *AVO-2* the least (12). Each sub-category was found in at least three raters' TAPs, which suggested that these sub-categories were common strategies employed by these raters.

Drawing on the coding scheme, the initial version of CERSQ was drafted by transforming 28 sub-categories into 28 items, among which three (Item 4, 13, and 27) are negatively worded and require reverse scoring. Raters need to judge each item on a five-point Likert scale from 1 to 5 (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree).

Psychometric Properties of CERSQ

Based on 367 valid responses to the initial version of CERSQ, an item analysis was performed in order to discover any problematic items. Five items which produced item-total correlations lower than 0.20 were deleted, resulting in a set of 23 items.

In order to estimate whether these variables in CERSQ could be identified, an EFA was conducted using a maximum likelihood extraction with a varimax rotation. KMO and Barlett's tests were conducted before EFA. It was found that KMO value (0.734) and result of Barlett's test (Chi-square = 858.95, df = 66, p < 0.001) made it appropriate to conduct EFA. The following criteria were adopted to determine the number of common factors to retain: (1) the eigenvalue >1; (2) the scree plot; (3) the amount of common variance explained; (4) the conceptual interpretability of the factor structure. It was found that the final version of the questionnaire contained 12 items, and the four-factor solution [*Self-evaluate (SEV)*, *Decide (DEC)*, *Diagnose (DIA)*, and *Compare (COM)*] could explain 59.92% of the total variance. It should be mentioned that *SEV* is a meta-cognitive strategy, whereas the other three cognitive. **Table 2** shows the factor loadings and factors with factor loads larger than 0.45 (Bastug, 2015).

TABLE 3 | Between factors relations in the final version of CERSQ.

	Factor 1 (*SEV*)	Factor 2 (*DEC*)	Factor 3 (*DIA*)	Factor 4 (*COM*)
Factor 1 (*SEV*)	1.00			
Factor 2 (*DEC*)	0.29**[a]	1.00		
Factor 3 (*DIA*)	0.21**	0.19**	1.00	
Factor 4 (*COM*)	0.41**	0.25**	0.18**	1.00

[a] ***Correlation is significant at the 0.01 level (2-tailed).*

In addition, correlations between four factors were also examined. **Table 3** shows the results.

It can be found that these four factors have positive and significant correlation with each other, which means that the increase in one factor would result in the rise of others. According to the standard of correlation strength (Muijs, 2004, p. 145), most factors correlated modestly with each other (r < 0.30) except that Factor 1 (*Self-evaluate*) and Factor 4 (*Compare*) is moderately correlated (r = 0.41), which provided evidence on the convergent and discriminant validity of CERSQ. Besides,

TABLE 4 | CERSQ's reliability in the initial and final version.

	Initial version	Final version
The questionnaire	0.75 (28 items)	0.73 (12 items)
MS	0.46 (6 items)	0.72 (3 items)
CS	0.70 (22 items)	0.65 (9 items)

these results provided quantitative evidence for the debated issue of the nature of *Compare* strategy. On the one hand, it was treated as a judgment strategy but with a self-control focus (Cumming, 1990) or a self-monitoring focus (Cumming et al., 2002). Besides, in Zhang's (2009) study of exploring CET4 essay raters' rating process and rater belief, it was considered a self-monitoring strategy. On the other hand, it was regarded as a sub-category of CS in Zhang's (2016) study on the relationship between raters' cognition with rating accuracy. Based on the above results, it would be more appropriate to label this strategy a cognitive strategy, although *Compare* is more closely correlated with *Self-evaluate* than other CS.

Cronbach's coefficient α was calculated to determine the internal consistency of the instrument, and a comparison was also made between the initial and final version of CERSQ. Results are shown in **Table 4**. Given to the small number of items in the final version, these values show an acceptable internal consistency for CERSQ.

DISCUSSION

In this study, a questionnaire aiming to measure CET4 essay rating strategy was developed and initially validated. Scale items were empirically derived from 14 experienced raters' verbal protocols while rating 10 CET4 essays. EFA was employed to determine the underlying structure of a number of variables. As a result, 16 items were deleted from the scale because of the low factor loading or conceptual vagueness. Consequently, the 12-item scale with four factors was developed, and each factor consists of three items. Further correlational and reliability analysis were performed to examine the psychometric properties of the instrument. Results exhibited acceptable internal consistency, and adequate convergent and discriminant validity.

In the first phase, there are some new findings about rating strategies compared with previous studies.

First, *Interpret* is the largest group of CS, and *INT-4* seems quite new in the literature to the best knowledge of the author. Raters not only read or reread the prompt or composition (Cumming et al., 2002), they were also found to explain the author's intention in Chinese as well. This strategy appeared in the lowly scored essays like Essay 3, 6, and 8. A possible reason may be that these essays are poorly written but have a certain length, reading the text directly may not be effective to create a text image, hence some raters would tend to use their mother tongue to explain the author's intention in order to understand the author.

Second, regarding the strategy of making inferences (*Inference*), it was found that raters not only inferred students' overall proficiency (*INF-1*) and their test strategies (*INF-2*) as documented in Cumming et al. (2002) and Zhang (2016), raters also inferred students' possible objective scores (*INF-3*). As some raters in Phase 2 were with rich experience of CET4 essay rating, they knew well how their performances would be evaluated. Therefore, they may consciously bring this knowledge to the rating process in order to maintain a good correlation coefficient. From the angle of Lumley's (2005) socio-cognitive model of rating process, it could be interpreted as a reaction to the institutional constraints imposed on raters (i.e., the quality control standard of raters' performances in the context of CET4 essay rating).

Third, *Self-evaluate* is the second largest category of strategies (with 209 frequencies). Raters tend to reflect on the severity/leniency of scores (*SEV-2*, 88), the sufficiency of their rating rationales (*SEV-3*, 86), and the scoring confidence (*SEV-1*, 35). It can be argued that, some raters in Phase 2, who were with rich experience in CET4 essay rating, adopted a heavy use of self-reflexive strategies (namely MCS), which echoes findings in the previous studies (Cumming, 1990; Cumming et al., 2002; Zhang, 2016).

Finally, the avoiding strategy (*AVO*), as a kind of meta-cognitive strategy, triangulated Baker's (2012) finding that some raters with the avoidant decision-making style tended to avoid extreme scores (like *AVO-1* and *AVO-2*) or have a central tendency (like *AVO-3*). In comparison, raters in the present study were either more concerned with the institutional criterion of performance evaluation (e.g., R2 and R8), or developing certain safe and productive strategies over time (Bejar, 2012) in order to deal with the heavy workload as well as the quality control mechanism (e.g., R6). As raters are not working in a vacuum, various institutional constraints such as the criterion to evaluate raters' performances, the heavy workload, and the duration of rating session, would cause raters to develop new rating strategies. In a word, the CET4 essay rating context provides a particular perspective on rater cognition, where Lumley's (2005) model can shed light on such findings.

In the second phase, the four factors (*Self-evaluate*, *Decide*, *Diagnose*, and *Compare*) were found to be distinct but interrelated constructs constituting CET4 essay rating strategy, which could be explained by the following reasons.

First, to adopt the *Diagnose* strategy is an ideal choice for raters because it can help them complete the heavy workload efficiently and obtain a high correlation coefficient simultaneously. The strategy is related with language accuracy, which is one of the most salient and distinctive characteristics in the L2 learners' uneven development profile (Hamp-Lyons, 1991; Weigle, 2002). Therefore, raters would attach great attention to language errors in students' essays in the rating practice. Besides, as raters are aware that their performances are evaluated against the correlation coefficient, they may realize that the effective way to cope with the quality control criterion would be focusing on certain distinctive language features, rather than take many features into account for an essay. Although this reductionist approach to rating would threaten the validity of the given score, it is effective for raters to achieve a good correlation coefficient.

Second, the use of *Compare* strategy is raters' rational choice to determine the score level, which confirms the previous findings like Cumming et al. (2002) and Zhang (2016). In the present study, this strategy includes two sub-strategies: one requires raters' comparing their scores with the external reference (*COM-3*) and the other two with the internal reference (*COM-1* and *COM-2*). Like the strategy of *Diagnose*, raters' awareness of the institutional standard of quality control would help them strive to be consistent in order to attain a high correlation coefficient. In this case, using the *Compare* strategy is a reasonable choice.

Third, as raters have to make a scoring decision for an essay, employing the *Decide* strategy is undoubtedly obligatory for all raters. Besides, raters have been trained explicitly to acquire the strategy in the training session, so it becomes a required skill for all of them.

Last, the *Self-evaluate* strategy, as a kind of meta-cognition, can be seen as raters' internal control of rating quality. It was found that raters would not only monitor the confidence about the score (*SEV-1*) and reflect on the leniency/harshness of the score (*SEV-2*), but also evaluate the sufficiency of the evidence which could buttress the judgment (*SEV-3*). As several qualitative studies (Cumming, 1990; Huot, 1993; Wolfe, 1997; Cumming et al., 2002; Zhang, 2016) found that more experienced or proficient raters have used more self-monitoring strategies, a follow-up study could be conducted in an attempt to investigate whether there is any significant differences between differently experienced raters in the use of the *Self-evaluate* strategy.

CONCLUSION

The merit of this study lies in the development of a useful instrument for investigating the raters' use and selection of decision-making strategies by using the mixed method, which makes an attempt to tackle the challenge that using small samples in the previous research on rater cognition limits the generalizability of the findings (Myford, 2012). This instrument might be used for other large-scale essay scoring contexts in China such as The National Matriculation English Test (NMET) and The Graduate School Entrance English Examination (GSEEE) (Cheng and Curtis, 2010), both of which are also characterized with the heavy workload and the time limit in practical rating.

However, due to the insufficiency of sampled raters, this study failed to employ confirmatory factor analysis (CFA) to examine the factor structure of the questionnaire. Hence, further validation of the instrument is warranted. Another limitation of the study was the sole use of verbal protocols to elicit raters' thought processes. Of course, more behavior evidence should be collected in order to triangulate the findings derived from this self-report tool.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

Research Project of Philosophy and Social Science of Ministry of Education of China (No. 17YJC740102). Guangdong Provincial Education Science Project (No. GD14XWW21). Guangzhou Municipal Social Science Fund (No. 2016GZGJ70).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2018.00027/full#supplementary-material

REFERENCES

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, B. A. (2012). Individual differences in rater decision-making style: an exploratory mixed-methods study. *Lang. Assess. Q.* 9, 225–248. doi: 10.1080/15434303.2011.637262

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

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Lang. Test.* 28, 51–75. doi: 10.1177/0265532210376379

Bastug, M. (2015). Scale development study for prospective teachers of online reading strategies. *Anthropologist* 19, 101–109. doi: 10.1080/09720073.2015.11891644

Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Measur. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x

Brown, A., Iwashita, N., and McNamara, T. (2005). *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks*. Available online at: https://www.ets.org/research/policy_research_reports/publications/report/2005/hsiw doi: 10.1002/j.2333-8504.2005.tb01982.x

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Q.* 25, 587–603. doi: 10.2307/3587078

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL Speaking test. *Lang. Assess. Q.* 12, 262–282. doi: 10.1080/15434303.2015.1053134

Cheng, L., and Curtis, A. (2010). *English Language Assessment and the Chinese Learner*. New York, NY: Routledge.

Connor-Linton, J. (1995). Looking behind the curtain: what do L2 composition ratings really mean? *TESOL Q.* 29, 762–765. doi: 10.2307/3588174

Creswell, J. W., and Plano Clark, V. L. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educ. Measur. Issues Pract.* 31, 10–20. doi: 10.1111/j.1745-3992.2012.00239.x

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. New York, NY: HarperCollins.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Lang. Test.* 7, 31–51. doi: 10.1177/026553229000700104

Cumming, A., Kantor, R., and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework. *Modern Lang. J.* 86, 67–96. doi: 10.1111/1540-4781.00137

DeRemer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assess. Writing* 5, 7–29.

Ebel, R. L., and Frisbie, D. A. (1991). *Essentials of Educational Measurement, 5th Edn.* New Jersey, NJ: Prentice Hall.

Ericsson, K. A., and Simon, H. A. (1984). *Protocol Analysis*. Cambridge, MA: MIT Press.

Freedman, S. W., and Calfee, R. C. (1983). "Holistic assessment of writing: experimental design and cognitive theory," in *Research on Writing: Principles and Methods*, eds P. Mosenthal, L. Tamor, and S. A. Walmsley (New York, NY: Longman), 75–98.

Glaser, B. G., and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine de Gruyter.

Hamp-Lyons, L. (1991). "Scoring procedures for ESL contexts," in *Assessing Second Language Writing in Academic Contexts*, ed L. Hamp-Lyons (Norwood, NJ: Ablex Publishing Corporation), 241–276.

Huot, B. (1993). "The influence of holistic scoring procedures on reading and rating student essays," in *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*, eds M. M. Williamson and B. A. Huot (Cresskill, NJ: Hampton Press, Inc.), 206–236.

Jin, Y. (2011). Fundamental concerns in high-stakes language testing: the case of the College English Test. *Pan-Pacific Assoc. Appl. Linguist.* 15, 71–83.

Jin, Y. (2017). Construct and content in context: implications for language, teaching and assessment in China. *Lang. Test. Asia* 7, 1–18. doi: 10.1186/s40468-017-0032-5

Johnson, J. S., and Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Lang. Test.* 26, 485–505. doi: 10.1177/0265532209340186

Kane, M. T. (2006). "Validation," in *Educational Measurement*, Vol. 4, ed R. L. Brennan (Westport, CT: Praeger Publishers), 17–64.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: a mixed methods approach. *Lang. Test.* 26, 187–217. doi: 10.1177/0265532208101010

Knoch, U., and Chapelle, C. (2017). Validation of rating processes within an argument-based framework. *Lang. Test*. 34, 1–23. doi: 10.1177/0265532217710049

Kobayashi, H., and Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: cultural rhetorical pattern and readers' background. *Lang. Learn.* 46, 397–433.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Q.* 26, 81–112. doi: 10.2307/3587370

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Lang. Test.* 19, 246–276. doi: 10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing Second Language Writing*. Berlin: Peter Lang.

May, L. A. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. *Lang. Test.* 26, 397–421. doi: 10.1177/0265532209104668

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

Milanovic, M., Saville, N., and Shuhong, S. (1996). "A study of the decision-making behaviour of composition markers," in *Performance Testing, Cognition and Assessment*, eds M. Milanovic and N. Saville (Cambridge: Cambridge University Press), 92–111.

Ministry of Education (2006). *College English Curriculum Requirements.* Beijing: Foreign Language Teaching and Research Press.

Muijs, D. (2004). *Doing Quantitative Research in Education with SPSS*. London: Sage.

Myford, C. M. (2012). Rater cognition research: some possible directions for the future. *Educ. Measur. Issues Practi.* 31, 48–49. doi: 10.1111/j.1745-3992.2012.00243.x

O'Malley, J. M., and Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.

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

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Lang. Test.* 18, 303–325. doi: 10.1177/026553220101800303

Shohamy, E., Gordon, C. M., and Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Lang. J.* 76, 27–33.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educ. Measur. Issues Pract.* 31, 21–30. doi: 10.1111/j.1745-3992.2012.00240.x

Vaughan, C. (1991). "Holistic assessment: what goes on in the rater's mind?" in *Assessing Second Language Writing in Academic Contexts*, ed L. Hamp-Lyons (Norwood, NJ: Ablex Publishing Corporation), 111–125.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan Houndmills.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assess. Writing* 4, 83–106. doi: 10.1016/S1075-2935(97)80006-2

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *J. Writing Assess.* 2, 37–56.

Wolfe, E. W., Kao, C. W., and Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Commun.* 15, 465–492. doi: 10.1177/0741088398015004002

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: a mixed-methods approach. *Lang. Test.* 31, 501–527. doi: 10.1177/0265532214536171

Yang, H., and Weir, C. J. (1998). *The CET Validation Study*. Shanghai: Shanghai Foreign Language Education Press.

Zhang, J. (2009). *Exploring Rating Process and Rater Belief: Seeking the Internal Account for Rater Variability.* Doctoral dissertation, Guangdong University of Foreign Studies, Guangzhou, China.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assess. Writing* 27, 37–53. doi: 10.1016/j.asw.2015.11.001

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.