Rater Cognition in L2 Speaking Assessment: A Review of the Literature

Qie Han[1]
Teachers College, Columbia University

ABSTRACT

This literature review attempts to survey representative studies within the context of L2 speaking assessment that have contributed to the conceptualization of rater cognition. Two types of studies are looked at: 1) studies that examine *how* raters differ (and sometimes agree) in their cognitive processes and rating behaviors, in terms of their focus and feature attention, their approaches to scoring, and their treatment of the scoring criteria and non-criteria relevant aspects and features of the speaking performance; 2) studies that explore *why* raters differ, through the analysis of the interactions between several rater background factors (i.e., rater language background, rater experience and rater training) and their rating behaviors and decision-making processes. The two types of studies have improved our understanding of the nature and the causes of rater variability in their perception and evaluation of L2 speech. However, very few of those studies has drawn on existing theories of human information processing and research on strategy use, which can explain on a cognitive-processing (Purpura, 2014) level what goes on in raters' mind during assessment. It is argued as a final conclusion that only based on established frameworks of human information processing and research on (meta)cognitive strategy use can rater cognition be explored with more depth and breadth.

INTRODUCTION

Human raters are normally involved in the evaluation of verbal responses that examinees produce in L2 speaking assessment. They provide scores for a variety of high-stakes speaking assessments, and based on those scores, stakeholders make inferences about examinee's oral English proficiency regarding various opportunities including employment, education, and immigration. Raters as human beings, however, are susceptible to biased, inaccurate, and inconsistent patterns of judgment, which have been defined in educational assessment as *rater effects*. According to Wolfe and McVay (2012), rater effects are "patterns of ratings that contain measurement errors" and can thus lead to issues regarding validity in human scores (p. 32). One of the most commonly studied rater effects is rater severity/leniency, which reflects a rater's systematically harsher or more lenient than accurate rating patterns. In a speaking test, a severe/lenient rater may assign consistently lower/higher scores to examinees whose actual

[1] Qie Han is an Ed.D. student in Applied Linguistics at Teachers College, Columbia University. Her current research interest is to study raters' cognitive processes and rating behaviors while assessing second language (L2) speaking performance. Correspondence should be sent to Qie Han, 316 Zankel Hall, 525 W 120th Street, New York, NY 10027. E-mail: qh2139@tc.columbia.edu

abilities are not commensurate with the levels of speaking performance depicted by the scores, therefore rendering the measurement invalid and biased. Other commonly studied rater effects include rater inaccuracy and differential rater functioning over time (DRIFT) (Wolfe & McVay, 2012). Despite their prevalent existence, rater effects are not always reflected in the actual scores assigned by raters (Orr, 2002). In a speaking test, two raters may give the same scores on the same criteria for the same speaking performance, whereas their perceptions and interpretations of the performance may diverge. Such latent variability in raters' cognitive processes has necessitated further examination on the validity of those test scores and the inferences that are made based on them. As a result, rater cognition has become an important area of inquiry in language assessment so that the meaning of and the inferences made based on human scores can be justified (Bejar, 2012; Kane, 2006). As automated scoring systems are gaining potential in rating constructed responses (Livingston, 2009), a validity argument for human scores can also provide derivative argument for the validity of automated scores, since automated scoring systems are often "trained on human scores used as the criterion to be predicted from features extracted from the responses" (Attali & Burstein, 2006, as cited in Bejar, 2012, p. 2). Essentially, rater cognition research can inform our understanding of the exact nature of rater variability and thus help us tackle practical problems regarding test score validation and rater training.

The purpose of this literature review is to scrutinize representative studies within the context of L2 speaking assessment that have contributed to the conceptualization of rater cognition. Since "rater cognition" has often been used as an umbrella term in the context of L2 assessment, what rater cognition entails as a definition is clarified at the beginning of the literature review. The clarification is based on influential theories and findings in language assessment, mainstream education and cognitive psychology. Subsequently, the current status of the research on rater cognition in L2 speaking assessment is briefly overviewed. The overview divides the existing research into two types of studies— *how* and *why* raters differ (and sometimes agree) in their cognitive processes and rating behaviors during the assessment L2 speaking performance. Following the overview, specific studies of each of the two types are reviewed and critiqued. In the end, a final summary of the two types of studies is provided, together with a discussion of the two major gaps found in the existing research. Finally, a tentative model of a hypothesized process of rating L2 speaking performance is proposed based on both the findings from this literature review and relevant theories and studies about the architecture of human information processing and (meta)cognitive strategy use.

It is worth mentioning, however, that even though this literature review is intended to examine studies on rater cognition in L2 *speaking* assessment only, research in writing assessment is related to sometimes as a frame of reference for conceptualizing rater cognition in speaking assessment. Because speaking and writing as two modalities of assessment are both categorized under the task type of constructed response in L2 assessment, raters' evaluation process of both modalities share some similarities. For example, they need to get trained, utilize a scoring rubric, focus on certain scoring criteria (e.g., content, grammar, organization) and adopt some commonly-used rating strategies (e.g., deciding, comparing, inferencing) to judge both types of responses. However, findings from writing assessment should not be transferred indiscriminately to the context of speaking assessment, because the two modalities differ in terms of both the language features to be judged as well as the ways in which raters interact with examinee responses (Davis, 2012). Therefore, based on the rationale above, this literature review will reference findings about rater cognition from writing assessment only occasionally and critically and as a contrast to the relative paucity of existing research in speaking assessment.

This literature review is essentially an examination of the research on rater cognition in the context of L2 speaking assessment.

Research on Rater Cognition in L2 Speaking Assessment

Definition of rater cognition in L2 assessment

So far, there have been very few explicit definitions of rater cognition in research on L2 speaking assessment. Davis (2012) mentioned rater cognition as "the mental processes occurring during scoring, at either a conscious or unconscious level" (p. 9). According to this statement, rater cognition mainly includes the various cognitive processes that raters go through during rating L2 speaking performance, either consciously (deliberately, analytically) or unconsciously (automatically, holistically). Although this definition entails the essential aspect of rater cognition, rater cognition is a much more complex matter and needs to be conceptualized based on empirical research. According to Bejar (2012), who has reviewed existing research on rater cognition in multiple assessment contexts, rater cognition has been examined from two major dimensions: "the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (p. 2). The attributes of the raters refer to various rater characteristics and other rater background factors that may influence raters' judgment process. These attributes mainly include raters' age, gender, occupation, educational background, language background (i.e., native/non-native speaking rater comparisons, matches between rater and examinee language background) and rater expertise (i.e., training and qualifications, ESL teaching experience, ESL rating experience). A thorough understanding of rater characteristics and other background factors is fundamental to an analysis of rater behavior or decision-making processes, because they serve to explain "why raters assign ratings the way they do and what attributes or elements they still need to improve their rating performance" (Kim, 2015, p. 241). Among all the attributes, rater language background (Johnson & Lim, 2009; Wei & Llosa, 2015; Zhang & Elder, 2011, 2014), rater experience (Barkaoui, 2010; Cumming, 1990; Davis, 2012, 2015; Isaac & Thompson, 2013; Kim, 2011, 2015) and rater training (Davis, 2012, 2015; Kim, 2011, 2015; Knoch, 2011; Xi & Mollaun, 2009; Weigle, 1998) have been most frequently examined for their effects on raters' cognitive processes and rating behaviors in L2 assessment.

Another important dimension of rater cognition concerns raters' mental processes that are invoked during rating. Raters' mental processes mainly pertain to the architecture of human information processing (Baddeley, 2012; Baddeley, Eysenck, and Anderson, 2009; Gagné, Yekovich, & Yekovich, 1993; Purpura, 2012), and the various (meta)cognitive strategies (Purpura, 2012) that raters deploy throughout rating. The architecture of human information processing can explain the underlying structure and processes (e.g., short-term, working and long-term memory) involved in the encoding, storage and retrieval of information during rating. In complement to that, the (meta)cognitive strategies (e.g., attention, reasoning, judgment, planning, monitoring) can be interfaced with that architecture, to explain in further detail what goes on in raters' minds during rating. So far, there have been very few models which have attempted to delineate the process of the rating of speaking and writing assessment by interfacing the architecture of human information processing with (meta)cognitive strategy use. In other words, very few researchers have taken a "cognitive-processing" (Dehn, 2008; Purpura, 2014) approach to conceptualize the process of the rating of speaking and writing assessment. Studies

taking a cognitive-processing approach consider "the components of the mind's cognitive architecture and the functions of these components (e.g., input processing, output processing, executive processing)" as integral parts of raters' scoring performance (Dehn, 2008, as cited in Purpura, 2014, p. 15). Among all the cognitive functions involved, (meta)cognitive strategy use are typically included as the "strategic aspects of cognitive functioning" (as cited in Purpura, 2014, p. 17). In L2 writing assessment, a moderate number of studies (Cumming, Kantor, & Powers, 2002; Milanovic, Saville, & Shuhong, 1996; Wolfe, 1997) attempted to postulate the process of and the relevant (meta)cognitive strategies involved in scoring L2 essays. However, none of them have drawn on models of human information processing as the basis of their framework, as Freedman & Calfee (1983) did for conceptualizing a cognitive framework for scoring L1 essays. In L2 speaking assessment, however, there has been almost no models which have attempted to explain the process of rating by interfacing the architecture of human information processing with the range of (meta)cognitive strategies that raters may deploy during rating. Less likely are there models which have incorporated rater background factors into a cognitive-processing framework of rating.

An overview: Two types of studies of rater cognition in L2 speaking assessment

Although rater cognition has been broadly explored in L2 writing assessment (e.g., Barkaoui, 2007, 2010; Cumming, 1990; Cumming, Kantor, & Powers, 2002; Freedman & Calfee, 1983; Lumley, 2002; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Smith, 2000; Wolfe, 1997), there is relatively limited research on rater cognition in L2 speaking assessment. Quantitatively, in L2 speaking assessment, rater severity, consistency and interaction with other aspects of rating have been studied by statistical analysis using Rasch (Bonk & Ockey, 2003; Brown, 1995; Eckes, 2005; Hiseh, 2011; McNamara, 1996; Orr, 2002). Qualitatively, rater cognition and rating process have been examined in only a limited number of isolated and exploratory studies in L2 speaking assessment. Relevant studies can be classified in two types.

The first type looked at *how* raters tend to differ or agree in their cognitive processes and rating behaviors. These studies looked at various aspects of rating L2 speaking performance. The most frequently explored aspects include: 1) rater focus and feature attention; 2) raters' approaches of rating; 3) raters' treatment of the scoring criteria and non-criteria relevant aspects of performances. However, almost none of those studies has taken a "cognitive processing approach" (Purpura, 2014) to examining the rater judgment process, i.e., to inspect the underlying processes and strategies invoked while raters are "attempting to understand response input, formulate a mental representation of the response, compare the response representation with that in the rubric, and evaluate the response in those terms" (p. 18). As a result, there has been no cognitive-processing models of the process of rating in L2 speaking performance.

The second type of qualitative studies attempt to explain *why* raters differ in their cognitive processes and rating behaviors (Davis, 2012, 2015; Kim, 2011, 2015; Kim, 2009; Isaacs & Thompson, 2013; Winke et al, 2011, 2012; Zhang & Elder, 2011, 2014), with an increasing attention to the influence of various rater characteristics and background factors, such as rater language background (i.e., rater L1 and L2), rater expertise and rater training. Yet again, the effects of some of the rater characteristics and background factors, such as rater experience and rater training, are far from being fully studied. Nor have those rater characteristics and background factors been associated with the cognitive mechanisms and mental processes of the raters. For example, what are the differences between raters with different characteristics and

background factors in their activation of various components of their architecture of information processing, as well as their uses of various (meta)cognitive strategies? So far, almost none of the studies in L2 speaking assessment has looked at the impacts of rater background from those angles. All in all, rater cognition research in L2 speaking assessment has yet to be grounded in a firm theoretical basis of human information processing, which can provide in-depth understanding of what goes on in raters' mind during scoring. Future research taking a cognitive-processing approach (Purpura, 2014) in L2 speaking assessment can provide us with more insight of the nature of rater cognition and rater variability.

Examination of specific studies on rater cognition in L2 speaking assessment

Type 1 Studies: An inquiry of raters' cognitive processes and rating behaviors. Recent studies on rater cognition in L2 speaking assessment looked at how raters tend to agree or differ in their cognitive processes and rating behaviors from different aspects. Those aspects mainly include raters' focus and feature attention, their approaches to scoring, and their treatment of the scoring criteria and non-criteria relevant aspects of performance. This is also the type of studies that most directly taps into raters' mental processes during scoring.

Rater focus and feature attention has been the most frequently examined facet of rater cognition in L2 speaking assessment. Some studies show that raters tend to focus on different aspects and features of a performance, and have different interpretations or apply different standards when judging the same performance (Ang-Aw & Goh, 2011; Orr, 2002). Orr (2002) examined retrospective verbal reports from 32 trained raters of the First Certificate in English (FCE), and found that raters did not heed the same aspects of the assessment criteria, and applied different standards when rating speaking performance. Likewise, Ang-Aw & Goh (2011) examined discrepancies in rater judgments via verbal protocol analysis and found that not only did raters place different emphases on factors of speaking performance assessed, but they also had different interpretations of candidates' performance and used different yardsticks when judging the performance.

On the contrary, some studies demonstrate that raters generally agree on what aspects and features of performance to be valued, although they may have different orientations to the same aspects or features (Brown, Iwashita, & McNamara, 2005). To scrutinize what raters attend to while judging performance on speaking tasks, Brown, Iwashita, & McNamara (2005) used retrospective verbal reports to explore raters' perceptions of the aspects and features of a speaking performance. They collected, categorized, and quantified the percentages of all the rater comments, and found out that the mostly commented categories include linguistic resources, phonology, fluency, and content. Within each category raters also considered a range of specific performance features. While there was a general agreement as to what aspects of performance they focused on, raters appeared to diverge in their judgment of the levels of proficiency indicated by some features. For instance, some raters considered the re-use of input vocabulary in test-takers' responses as an indication of a higher level of proficiency, whereas others believed test-takers of higher proficiency should be able to paraphrase or find alternatives for the input vocabulary. This finding corroborates that of Brown (2000), who found that raters of IELTS interviews held different opinions with regard to the oral proficiency levels indicated by test candidates' use of self-correction, circumlocution, and self-clarification strategies. Brown et al.'s (2005) findings have contributed to our understanding of rater focus and feature attention by providing a fine-grained analysis of what raters put emphasis on when judging performance on

speaking tasks. However, their study is concerned not with raters' mental processes of arriving at a judgment of proficiency, but simply with identifying the features on which raters focus when attempting to reach that judgment, thus leaving much more about rater cognition to be explored.

With the increasing importance of communicative competence, researchers in L2 speaking assessment have also investigated the categories and features of interactional competence that are salient to raters of paired speaking tests. Raters are found to differ in the extent to which they view paired L2 speaking interactions as co-constructed (May, 2006), although later research showed that there are some main categories and features of interactional competence that raters tend to agree on (Ducasse, 2010; Ducasse & Brown, 2009; May, 2011). May (2006) examined two trained and experienced raters' orientations while rating paired candidate discussion tasks through the use of stimulated verbal recall. In alignment with general findings about raters' focus and feature attention, she found that different dimensions of the rating scales appeared to be more salient to individual raters. One of those dimensions is the extent to which the speaking performance is co-constructed by multiple participants (e.g., test-taker and interviewer, or different test-takers), and raters differed in their acknowledgement of the degree of co-construction. That has led to further explorations of raters' orientations to the co-constructed nature of paired speaking tests. Through analyzing raters' verbal protocols, researchers (Ducasse, 2010; May, 2009, 2011) identified some of the main categories and features of interactional competence that raters put emphasis on, such as non-verbal interpersonal communication (e.g., gestures, gaze, laughter), interactional listening comprehension, and interactional management (e.g., topic change and turn organization) (Ducasse, 2010). Not only do those identified features have implications for redefining the construct of oral proficiency in interactional speaking contexts and operationalizing this in rating scales, they also enrich our understanding of rater cognition by demonstrating the variations in raters' focus and feature attention while scoring a different L2 speaking task type.

Apart from rater focus and feature attention, most studies in L2 speaking assessment have investigated raters' judgment process and demonstrated that raters adopt different approaches to evaluating speaking performance. Researchers in L2 speaking assessment generally concur with each other on the decision-making approaches or styles they have discovered. Pollitt and Murray (1996) examined raters' decision-making process while comparing and contrasting pairs of candidate's performance during Certificate of Proficiency in English (CPE) oral interviews. They found that raters adopted two contrastive approaches of assessment, i.e., a synthetic (intuitive) approach which allowed them to generate holistic impression of examinee's performance, versus an analytical approach through which raters add up the scores for each utterance of candidates' performance. Such finding has been corroborated by a few other studies on rater cognition in L2 speaking assessment (Brown, 2000; May, 2006), which basically agree with the dichotomy of the synthetic versus the analytical approach of rating. Based on those findings, Ang-Aw & Goh (2011) explored discrepancies in rater judgments via the analysis of rater verbal protocols and discovered a third approach to rating— a "mixed" approach which combines the two rating approaches. Even though findings about rating approaches in L2 speaking assessment seem to be quite congruent, none of them to my knowledge has been associated with theoretical explanations from human information processing research. In other words, those findings would be more convincing if evidence from human cognitive processing research were provided to support the dichotomy of synthetic (intuitive) versus analytical rating approaches identified, as has been done by researchers of assessment in the context of examination marking (Greatorex & Suto, 2006; Suto, Crisp, & Greatorex, 2008; Suto & Greatorex, 2006; Suto & Greatorex, 2008).

They categorized raters' approaches to scoring based on the dual processing theory (Kahneman & Frederick, 2002; Stanovich & West, 2002), which distinguished human cognitive operations into two qualitatively distinct but simultaneously active systems. System 1 represents the quick, automatic, skilled and intuitive thought processes, which occur in parallel with the slow, reflective, effortful and rule-governed System 2 thought processes. It is likely that the two contrastive approaches to rating (intuitive vs. analytical) discovered in L2 speaking assessment correspond to the dichotomy and co-existence of the two qualitatively different systems of human cognitive operations. It would be more interesting to associate the use of disparate rating approaches and systems of thought processes with raters' experience in scoring, which might serve to explain, according to some findings in existing research on writing assessment (Sakyi, 2003; Wolfe, 1995, 1997, 2006; Wolfe, Kao, & Ranney, 1998), why more proficient raters tend to score faster, rely less on the scoring rubric, and adopt a more holistic approach to rating. There has also been very little explanation as to what happens on a cognitive-processing level (Purpura, 2014) in the rater's mind when they employ or switch between different rating approaches. Explorations in greater details about the relevant components of the internal structure of mind and the use of corresponding (meta)cognitive strategy would provide a clearer picture of raters' mental actions and thought processes during the intricate process of rating L2 speech.

Researchers in L2 speaking assessment also seem to be in agreement with regard to raters' treatment of the scoring criteria and non-criteria relevant aspects and features of the performance. Brown (2000) used stimulated verbal recall to tap into eight trained raters' decision making within the context of IELTS interviews, and found that raters focused not only on criteria specified in the scales, but also on aspects that were not explicitly included in the scales, such as the quality/maturity of the examinees' ideas. This finding was corroborated by Meiron (1998), who examined rater behavior on a picture narrative task, and found that in addition to the specified criteria, raters also reported using certain "self-generated features not mentioned in the scoring rubric" (as cited in Brown, Iwashita, McNamara, 2005, p. 6). In correspondence with previous studies, Orr (2002) used retrospective verbal reports from 32 trained raters to help interpret test scores on the First Certificate in English (FCE). He found that raters did not heed the same aspects of the assessment criteria, and heeded a wide range of non-criterion relevant information, such as candidates' presentation of him/herself, and their communicative success and failure. May (2006) examined two trained and experienced raters' orientations while rating paired candidate discussion tasks through the use of stimulated verbal recall. As she found out, apart from the features and aspects of the performance that are included in the rating criteria, raters appeared to have "fleshed out the criteria in the band descriptors with features that were not explicitly mentioned in the band descriptors" (p. 13). For example, for the criteria of the range of language use, one rater valued the use of idiomatic and "natural" language, whereas the other rater did not mention this feature in the verbal protocols. Besides, more than 30% of rater comments in her study alluded to non-criterion aspects of the performance, such as their first impression of candidates, the confidence level and the sense of humor of candidates, the quality, complexity, relevance and logic of candidates' ideas. All those studies have verified that not only do different categories of the rating scales appear to be more salient to each rater, many non-criteria relevant aspects and features of performance are also taken into account for evaluation. One major drawback of this group of studies, though, is that they did not further explain what happened on a cognitive-processing level when raters heeded aspects not indicated in the rating criteria. For example, what past personal, professional or cultural knowledge and experiences might have triggered raters' attention to certain non-criteria relevant features of the responses,

and where do those knowledge and experiences fit in the architecture of human information processing? One possible explanation is that raters may have formed a prior understanding of what composes the construct of oral proficiency through their past personal, professional or cultural experiences, and have stored such knowledge and experiences in their long-term memory. Despite rater training, the mental rubric representations they have formed may still be susceptible to the influences of their prior knowledge and experiences, which they can retrieve from their long-term memory to guide their judgment process in combination with the assessment criteria. The whole series of cognitive operations should take place in their working memory, under the control of the central executive component (Baddeley, Eysenck, & Anderson, 2009). It can be clearly seen that only through the lens of human information processing can we elucidate the deeper, underlying cognitive processes and elements that might have influenced raters' attention and judgment process. Thus, a cognitive-processing approach should be adopted in future research to enhance our understanding of the process of rating.

To summarize, researchers in L2 speaking assessment examined rater judgment process from three major aspects, and generally reached similar conclusions. Fine-grained analyses on raters' selective attention to and disparate perceptions of the aspects and features of speaking performance have provided a relatively comprehensive picture of what raters tend to heed and how they may diverge in their evaluations. Raters' approaches to scoring and decision-making have also been investigated to show the distinct systems of thought processes that raters follow to reach their conclusions. Raters' treatment of the scoring criteria and non-criteria relevant aspects and features of the performance have also been scrutinized to provide information on how raters internalize and form a mental representation of the scoring criteria, and how they manipulate that mental representation to adapt to different scoring situations instead of adhering strictly to the marking scheme. However, with all those results and conclusions, researchers in L2 speaking assessment mostly examined rater cognition in a limited number of empirical, explorative studies. Existing findings about the rating process of L2 speaking performance are mostly descriptive instead of cognitive-processing oriented. Rater cognition research has yet to be grounded in a firm theoretical basis of human information processing, which can provide an in-depth understanding of what goes on in raters' minds during scoring.

Type 2 Studies: The influence of rater characteristics and other rater background factors. In complement to the studies which looked at *how* raters differ, studies on the effects of rater background factors attempt to explain *why* raters differ, with an increasing attention to the effects of rater language background, rater expertise and rater training on raters' cognitive processes and rating behaviors. Findings from both types of studies can be combined to provide a useful frame of reference for conceptualizing rater cognition in future research.

Rater language background (i.e., native/non-native speaking rater comparisons, matches between rater and examinee language background) has received major attention among researchers in L2 speaking assessment. A representative study that examined the cognitive differences between native and non-native speaking groups of raters was conducted by Zhang & Elder (2011, 2014), who investigated ESL/EFL teachers' evaluation and interpretation of oral English proficiency in the national College English Test-Spoken English Test (CET-SET) of China. They found that NS raters attended to a wider range of abilities when judging candidates' oral test performance than NNS raters. NS raters also tended to emphasize features of interaction while NNS raters were more likely to focus on linguistic resources such as accuracy. Similarly, Gui (2012) investigated whether American and Chinese EFL teachers differed in their

evaluations of student oral performance in an undergraduate speech competition in China. He found that the American raters provided more specific and elaborated qualitative comments than the Chinese raters. The raters also differed in their judgment of students' pronunciation, language usage, and speech delivery. One unique difference was related to raters' comments on students' nonverbal communication skills. The Chinese raters provided mostly positive comments about the gestures and other non-verbal demeanors of the students as a group, while the American raters were mostly critical. Both Zhang & Elder's (2011, 2014) and Gui's (2012) studies have offered some interesting revelations as to the differences in the perception of oral English proficiency and the pedagogical priorities between these two groups of raters. However, they seem to mainly focus on the aspects and features of language performance raters heed, leaving other important aspects of rater cognition, such as raters' decision-making behaviors and rating approaches, not thoroughly attended to. Another set of limitations also exist with regard to the validity and the generalizability of these results. The first limitation lies in the homogeneity of the student samples selected in both studies. Chinese students who share the same L1 and similar educational background might undermine the generalizability of the results to other test-taker populations. There is also limitation with regard to the validity of using written comments as the major data for analysis, which might not offer a full account of raters' in-depth rating behaviors. The last impediment to the validity of the results from both studies, as had been discussed in precedent studies on the influence of rater language background (Brown, 1995; Kim, 2009), pertains to the possibility that variables other than rater language background, such as raters' scoring experiences or their places of residence, could have caused the variance in ratings instead. Rater language background thus ended up in the original results as a proxy variable. This limitation has raised the question of whether language background is "a particularly meaningful category as far as predicting raters' behavior is concerned" (Zhang & Elder, 2014, p. 320).

Another type of research on rater language background attempted to find out whether raters tend to bias in favor of test-takers whose language backgrounds are related to theirs. Researchers have looked at the influence of both rater L1 and rater L2 and seem to diverge in their opinions. Winke, Gass, & Myford (2011, 2012) investigated whether raters were influenced by the link between their L2 and test-takers' L1 through scoring the TOEFL iBT speaking test. Both statistical results and qualitative data analyses suggested that raters tended to assign scores that were significantly higher than expected to test takers whose L1 matches their L2 (i.e., heritage status), due to familiarity and positive personal reactions to test-takers' accents and L1. On the contrary, Wei & Llosa (2015) examined the differences between American and Indian raters in their scores and scoring processes while rating Indian test takers' responses to the TOEFL iBT speaking tasks. They found no statistically significant differences between Indian and American raters in their use of the scoring criteria, their attitudes toward Indian English, or in the internal consistency and severity of the scores. In-depth qualitative analysis revealed that some Indian raters even held negative attitudes toward Indian English, due to factors more complicated than their own language background. For example, the negative judgments one rater received about his native language caused him to believe that adopting standard American English is important for surviving in the United States. As a result, this rater might not have endorsed test-takers' shared language background. The findings of this study suggest that sharing a common language background does not guarantee a positive evaluation of test-takers' L2 speaking performance after all. However, issues regarding the small and homogeneous sample of Indian raters used might undermine the generalizability of the findings of this study, which should be further examined by including raters and test-takers of other language varieties.

So far in L2 speaking assessment, researchers have provided statistical and qualitative support for various hypotheses regarding whether raters are potentially biased toward test-takers from a similar language background. However, they have yet to examine whether deeper, underlying cognitive differences exist in raters' scoring processes, such as their approaches to rating and their focus and feature attention, while they are evaluating the performance of test-takers with mixed language backgrounds. One of the studies that attempted to tap into those cognitive differences was conducted by Xi & Mollaun (2009, 2011), who investigated the extent to which a special training package can help raters from India to score examinees with mixed first language (L1) backgrounds more accurately and consistently. As they found out, the special training not only improved Indian raters' consistency in scoring both Indian and non-Indian examinees, but also boosted their confidence in scoring. Those findings led to further discussion of whether raters adopted different styles of rating depending on the match between their and the examinees' first languages. For example, after the special training, the raters from India may have employed more analytical approaches to scoring Indian examinees while engaging in more impressionistic, intuitive evaluations for examinees whose L1s were not familiar to them (Xi & Mollaun, 2009), thus balancing out their tendency to bias toward test-takers of their own language background. However, the researchers could only make hypotheses about the change in raters' cognitive styles due to lack of direct empirical evidence (e.g., raters' verbal protocol data), which could have served to corroborate their quantitative findings.

Apart from rater language background, rater experience and rater training are also important factors that are found to affect raters' rating styles and behaviors in L2 speaking assessment. Among the series of studies that have explicitly examined the effects of experience on raters' cognitive processes and rating behaviors in language testing, the vast majority were conducted in writing assessment (Barkaoui, 2010; Cumming, 1990; Delaruelle, 1997; Lim, 2011; Myford, Marr, and Linacre, 1996; Sakyi, 2003; Wolfe, 1997, 2006; Wolfe, Kao, & Ranney, 1998). Research findings in writing assessment generally seem to agree that prior teaching or testing experience influences raters' decision making processes (Davis, 2012). Experienced raters are found to score faster (Sakyi, 2003), consider a wider variety of language features (Cumming, 1990; Kim, 2011; Sakyi, 2003), and are more inclined to withhold premature judgments in order to glean more information (Barkaoui, 2010; Wolfe, 1997). In terms of rater training, the majority of the studies in both writing and speaking assessment seems to suggest that training does not completely eliminate the variability existing in either rater severity (Brown, 1995; Lumley & McNamara, 1995; Myford & Wolfe, 2000) or their scoring standards and decision making processes (Meiron, 1998; Orr, 2002; Papajohn, 2002; Winke, Gass & Myford, 2011).

In contrast to the relatively larger number of studies on rater experience and rater training in L2 writing assessment, researchers in L2 speaking assessment have only recently begun to examine the impacts of those two rater background factors on raters' scoring processes and behaviors (Davis, 2012, 2015; Isaacs & Thompson, 2013; Kim, 2011, 2015). Kim (2015) compared rater behaviors across three rater groups (novice, developing, and expert) in the evaluation of ESL learners' oral responses, and examined the development of rating performance within each group over time. The analysis revealed that the three groups of raters demonstrated distinct levels of rating ability and different paces of progress in their rating performance. Based on her findings, she concluded that rater characteristics should be examined extensively to improve the current understanding of raters' different needs for training and rating. She also discussed her own conceptualization of rater characteristics and relative expertise drawing on relevant literature in writing assessment (e.g., Cumming, 1990; Delaruelle, 1997; Erdosy, 2004;

Lumley, 2005; Sakyi, 2003; Weigle, 1998; Wolfe, 2006), and proposed perhaps the most up-to-date framework of rating L2 speaking performance germane to those rater characteristics. According to Kim (2011, 2015), rater expertise is composed of four concrete rater background variables (i.e., experience in rating, Teaching English to Speakers of Other Languages [TESOL] experience, rater training, and coursework). The interactions of those rater background variables influence the rating-related knowledge and strategic competence that raters utilize during scoring, also known as their rating ability. Rating performance is then accomplished by raters harnessing their rating ability in an actual rating occasion. Kim's model is perhaps the most complicated framework of rating performance germane to rater background variables to date.

In another representative study on rater expertise in L2 speaking assessment, Davis (2012, 2015) investigated how raters of different rating proficiency scored responses from the TOEFL iBT speaking test differently prior to and following training. Considerable individual variations were seen in the frequency with which the exemplars were used and reviewed by raters, the language features mentioned during rating, and the styles of commenting by each rater (e.g., the array of topics covered and the amount of detailed explanation on specific points). The effects of training were reflected in the ways that raters gave more explicit attention to their scoring processes, and that they made fewer disorganized, or unclear comments over time. Both Kim's (2011, 2015) and Davis' (2012, 2015) research is comprehensive in terms of the rater background factors (i.e., rater experience interacting with training) they focused on and the research design and methods (i.e., mixed-method research design) they used to tap into the influence of those background factors. However, the data reported in their research primarily address raters' accuracy of interpreting the rating scales and performance level descriptors (Kim, 2015), and raters' conscious attention to specific language features while scoring (Davis, 2012), leaving other important aspects of rater cognition, such as the mental actions raters take to reach a scoring decision, not thoroughly attended to.

As a further attempt to investigate the cognitive differences between more and less experienced raters, Isaacs & Thompson (2013) examined the effects of rater experience on their judgments of L2 speech, especially regarding pronunciation. This study has discovered some fresh cognitive differences between experienced and novice raters, in terms of the (meta)cognitive strategies they use to harness their relative experience with ESL learners, their emotional reactions and attitudes toward their levels of experience, their rating focus and feature attention, their professional knowledge and TESOL vocabulary to describe L2 speech, and the relative lengths and styles of their verbal comments. Evidence from verbal protocols and post-task interviews suggested that experienced and novice raters adopted strategies to either draw on or balance out their perceived experience with L2 speech during scoring. For example, some experienced raters reported that they might have been affected by their experience with ESL learners in their comprehension and evaluation of learners' speech in comparison to non-ESL teachers. To neutralize the influence, some even attempted to envision themselves as non-ESL trained interlocutors when assigning scores. Conversely, several novice raters expressed feelings of inadequacy to be judges due to their insufficient experience specifying and assessing learner speech. In terms of rating focus and feature attention, experienced raters were more likely to identify specific pronunciation errors through either detailed characterization or imitation/correction of student speech. Compared to their novice counterparts, they also had a more flexible range of professional knowledge of L2 pronunciation and assessment, whereas novice raters were more uniformly lacking in their command of TESOL vocabulary to the extent that they had to think of more creative terms to describe L2 speech. Experienced raters were also

found to produce longer think-aloud and interview comments, since they almost unexceptionally provided anecdotal descriptions about their teaching or assessment practices. Even though the study attempted to gather evidence that shows raters diverged cognitively depending on their levels of rating experience, it is still unclear if the cognitive differences discovered were the essential ones that distinguish experienced raters from the novice ones. For example, it has not been verified if novice raters failed to articulate their perceptions of the speech due to their inadequate access to the vocabulary used by experienced raters, or rather due to the fact that experienced and novice raters were heeding qualitatively different dimensions of the speech overall, having different perceptions and interpretation of the construct and the scoring rubric, or following different approaches of rating. Therefore, it is important to examine in greater detail the factors that might have affected those raters' judgment process while scoring.

The most commonly studied rater background factors in L2 speaking assessment so far are rater language background, rater experience and rater training. What has been little known, however, is whether other sources of rater variability, for example, those related to the difference in raters' cognitive abilities, also affect raters' evaluation of L2 speaking performance. In a pioneering study, Isaacs & Trofimovich (2011) investigated how raters' judgments of L2 speech were associated with individual differences in their phonological memory, attention control, and musical ability. Results showed that raters who specialized in music assigned significantly lower scores than non-music majors for non-native like accents, particularly for low ability L2 speakers. However, the ratings were not significantly influenced by the variability in raters' phonological memory and attention control. Reassuring as it is that phonological memory and attention control are not found to induce bias in raters' assessments of L2 speech, this study is an initial attempt to tap into raters' cognitive abilities in relation to L2 speaking assessment, and calls for further explorations of the nature of the impacts of those abilities. One major caveat that might undermine the validity of the results here, as the researchers (Isaacs & Trofimovich, 2011) themselves have pointed out, is that phonological memory and attention control might not be as relevant to raters' perceptual judgments of L2 speech as alternative measures such as acoustic memory and the scope of attention, which raters might have drawn on more heavily to process and evaluate L2 speech (pp. 132- 133). Apart from that, the cognitive tasks used to measure raters' phonological memory (i.e., a serial non-word recognition task) and attention control (i.e., the trail-making test) might not be as effective as other tasks (e.g., nonword repetition or recall tasks) to yield the maximum association between those cognitive capacities and raters' perceptual judgments of L2 speech (Isaacs & Trofimovich, 2011, p. 132). The trail-making task, for example, was used to measure attention control of listeners who evaluate language performance. However, since the nature of the task is language neutral (p. 122), it does not seem to have much connection with real-life language processing and therefore, might not be the optimal measure of attentional control in the context of this study. In terms of the methods for data analyses, apart from preliminary statistical analyses of the results of cognitive ability measures, the study could also have benefited from collection and analyses of qualitative data (e.g., raters' verbal protocols and interview/questionnaire results) to capture more direct evidence of the effects of raters' cognitive abilities on their rating process. This study is obviously groundbreaking in terms of its implications to investigate rater cognition in relation to the architecture of human information processing and the functionality of the brain for L2 speaking assessments. However, apart from phonological memory and attentional control, the effects of many other cognitive abilities and mechanisms should also have been taken into account, such as raters' attention and perception, long-term memory (i.e., declarative, procedural and episodic

memory which might influence raters' mental representations of both the rubric and the L2 speech, and their rating styles and strategies), or reasoning and decision-making skills, to provide a more comprehensive picture of the important role that each component of the human cognitive architecture plays in the process of rating L2 speech. Musical ability, the factor that appeared to influence raters' judgments of accentedness in this study, needs to be explored in greater detail to explain how individual differences in musical expertise may impact rater behavior more precisely. Not only can drawbacks be found regarding the types of cognitive abilities explored in this study and the tasks used to measure them, how those cognitive abilities might affect the evaluation of a construct of speaking ability more broadly defined is also left unexplored (p. 136). For instance, researchers of this study only focused on three components of the speaking ability construct (i.e., accentedness, comprehensibility and fluency), without incorporating other elements (e.g. grammar and vocabulary), therefore largely diminishing the generalizability of the results to a wider variety of speaking tasks and oral proficiency constructs. The relatively homogenous sample of raters recruited (i.e., college majors who are untrained and inexperienced for scoring L2 speech) can also limit the generalizability of the results.

To summarize, by examining the interactions between various rater background factors and raters' judgment processes, researchers reached generally similar conclusions about the possible effects of different rater background factors on raters' cognitive processes and rating behaviors. Rater language background is found to be likely to affect the raters' focus and perception of oral proficiency when they are identified as native/non-native speaking individuals. Matches in language background between raters and examinees can also influence raters' comprehension and evaluation of examinees' interlanguage speech. Rater experience and rater training are also found to have impacts on raters' scoring approaches and styles, their commenting styles, their decision-making behaviors and strategy use, their focus and attention to performance features, and their interpretation and utilization of the scoring criteria. One of the groundbreaking studies (Isaacs & Trofimovich, 2011) attempted to look into the effects of individual differences in raters' cognitive abilities on their rating patterns and scoring process, but the results are not as convincing as expected due to a number of limitations. One major limitation among most of the studies is that they only focused on one or two isolated aspects (e.g. rater focus and feature attention) while exploring how rater background factors affect rating behaviors and cognitive processes, leaving other aspects not thoroughly explored, especially those that are directly related to raters' cognitive processes (e.g., raters' internal processing of information and their strategy use). Therefore, future research can improve our understanding of how various rater background factors might impact raters' judgment process by systematically exploring those influences from a cognitive-processing perspective.

DISCUSSION AND CONCLUSION

Human raters are usually engaged with the judgment of interlanguage speech that examinees produce in L2 speaking assessment. As a result, rater cognition has been extensively explored to inform our understanding of the exact nature of rater variability and help us tackle practical problems regarding score validation and rater training. As the above review has shown, existing studies in L2 speaking assessment which have contributed to the conceptualization of rater cognition can be categorized into two types: studies that examine *how* raters differ (and sometimes agree) in their cognitive processes and rating behaviors, and studies that explore *why*

they differ. The first type looked at how raters tend to differ or agree in their cognitive processes and rating behaviors, mainly in terms of their focus and feature attention, their approaches to scoring, and their treatment of the scoring criteria and non-criteria relevant aspects and features of performance. This is also the type of studies that most directly describes raters' mental processes during scoring. The second type attempted to explain why raters differ (and usually they do), through the analysis of the interactions between various rater background factors and raters' scoring behaviors. Regardless of disagreement in their findings, many researchers would probably argue that rater background variables, mainly composed of their language background, rating experience and training experience, can lead to individual variability and/or overtime adjustment in their judgment process when scoring L2 speech.

Two major gaps in existing research on rater cognition in L2 speaking assessment

Gap 1: A model of the rating process for speaking assessment is needed

In contrast to the variety of models of the rating process proposed by researchers in the assessment of writing (e.g., Cumming, Kantor, & Powers, 2002; Freedman & Calfee, 1983; Lumley, 2002; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Wolfe, 1997), there is almost no model of the rating process for speaking assessment. A natural question is whether those models of rating for writing assessment are transferrable to the context of speaking assessment. As Davis (2012) pointed out, this "transfer" approach is somewhat problematic because "the judgment of written and oral language performance represents very different cognitive and practical challenges, so models of judgment developed within writing assessment cannot be assumed to apply to the judgment of speaking as well" (p. 69). Some studies on speaking assessment have suggested that the rating process of spoken performance is distinct from, if not more complicated than, that of written performance, due to disparate cognitive processes (e.g., auditory processing), linguistic factors (e.g., familiarity with the speaker's L1 and accent) and contextual influences (e.g., independent vs. interactional oral performance). Especially with regard to the increasing emphasis on communicative competence, a complete set of features of interaction that seems to be exclusive for speaking assessment should be incorporated into the rating scales (Ducasse, 2010; May, 2011), and as a result, will influence each and every step of the rating process.

However, it should not be ignored that models for the assessment of performance in other modalities also represent how researchers have been gestating the scoring process in language assessment as a whole. Those models may in fact provide a useful frame of reference for conceptualizing raters' scoring processes in speaking tests. For example, Bejar (2012) has postulated one of the most advanced and comprehensive descriptive frameworks in the assessment of constructed responses, although it seems to apply to the assessment of written rather than spoken constructed responses. He delineated the rating process into two major phases: 1) The assessment design phase, during which raters read and form a mental representation of the rubric; 2) The actual scoring phase, where they read a response, form a mental representation of it, compare and contrast both mental representations and decide on a final score. Throughout the whole rating process, raters' judgments are susceptible to the impact of various background factors despite rater training in the assessment design phase. Although no model such as this has been postulated in the context of L2 speaking assessment, it is quite

possible that the rating process of L2 speech comprises of all these stages. For example, raters of L2 speaking performance also need to familiarize themselves with the scoring criteria (i.e., forming a mental rubric response) and the spoken responses (i.e., forming a mental rubric response). Instead of going back and forth to check what they hear against what is on the rubric, experienced raters may also be better at internalizing the scoring criteria so that they can evaluate the spoken performance holistically and intuitively, as has been suggested by relevant studies in L1 speaking assessment (Joe, 2008; Joe, Harmes, & Hickerson, 2011). Less experienced raters, on the other hand, may be more reliant on the scoring rubric and thus become more involved in analytical approaches to rating (Joe, Harmes, & Hickerson, 2011). However, they are also found to make progress on their understanding and application of the scoring criteria through extended training (Kim, 2015). Although raters hardly refer to the whole scoring rubric, they do attend to a selective range of the scoring criteria and non-criteria relevant features in the performance (Brown, 2000; May, 2006; Joe, Harmes, & Hickerson, 2011; Orr, 2002), which seems to reflect their cognitive processes of comparing their mental rubric representations with their mental response representations. Throughout the whole process of assessment, raters are found to be susceptible to the influences of their personal frames of reference, their professional and language background, and their personal reactions/connections to the spoken performance and the speaker (Brown, 2000; Brown, Iwashita, & McNamara, 2005; Joe, Harmes, & Hickerson, 2011; Kim, 2015; Wei & Llosa, 2015; Winke, Gass, & Myford, 2012).

Based on the literature review, directions of future research on the assessment of spoken performance can be pursued in alignment with Bejar's (2012) model. Research needs to be done to examine how raters of L2 speaking assessment form a mental rubric representation, which is related to the following observations in the literature: 1) raters compare speakers to exemplars or against each other (i.e., a type of rangefinders and benchmarks) during rating; 2) raters have dissimilar perceptions of the components of the oral proficiency construct, and place weighted emphasis on a limited set of aspects and features instead of consulting the full scoring rubric; 3) raters are inclined to focus on certain self-generated features that are not explicitly included or explained in the scoring rubric, and so on. How raters form a mental response representation and compare it with the rubric representation can also be explored with respect to findings such as: 1) they tend to disagree on their interpretations of a certain performance feature as an indication of oral proficiency; 2) they do not focus on the same aspects of the scoring criteria in the spoken performance, and diverge in how much non-criterion information they pay attention to; 3) they use distinguished approaches (i.e., holistic, analytical or mixed) to rating, and their choices or approaches might be related to their rating experience; 4) they sometimes draw on inferences about a candidate's personality, maturity, world knowledge, and so on, to justify their scoring patterns and decisions (Brown, 2000) ; 5) they incorporate their personal preferences into their decision-making, especially with respect to their language background and personal attitudes toward the response or the speaker ; 6) they differ in their construal of co-constructed spoken performance and their decision to assign separate scores (or a common score) to both conversation partners. To summarize, a variety of findings or questions in existing L2 speaking assessment research can be further investigated in light of Bejar's (2012) model for the assessment of constructed responses, so that all relevant findings can be reorganized and systematically incorporated into a framework of rating L2 spoken responses.

Gap 2: Research on rater cognition in L2 speaking assessment needs to be based on cognitive-processing theories

Another major gap, as has been discussed previously, relates to the fact that almost no studies in L2 speaking assessment has drawn on theories and empirical research about human information processing and strategy use. Due to the lack of empirical support here, there have been no models in L2 assessment that have attempted to interface the architecture of human information processing, the various (meta)cognitive strategies that raters deploy during rating, with a comprehensive descriptive framework of the rating process itself. As Purpura (2014) has pointed out, investigating rater cognition through a cognitive-processing approach (Purpura, 2014) will lead to more answers and solutions to the long-held concerns and interests by researchers regarding the validity of scores on performance tasks. In main-stream education, since three decades ago, researchers have proposed frameworks on learning and assessment (Bejar, 2012; Freedman & Calfee, 1983; Gagné, Yekovich, & Yekovich, 1993; Mayer, 2005) based on theories about the architecture of human information processing. In contrast, explorations of how the architecture of human information processing and the various (meta)cognitive strategies are involved in L2 processing began later in the field of L2 learning and assessment (Andringa, Hulstijn, Schoonen, van Beuningen, & Olsthoorn, 2012; Cohen & Upton, 2007; Hulstijn, Van Gelderen, & Schoonen, 2009; Phakiti, 2003a, 2003b, 2007; Purpura, 1997, 1998, 2012, 2014; Van Gelderen et al., 2004). Perhaps the only model of information processing as a basis for understanding the cognitive mechanisms underlying L2 performance was proposed recently by Purpura (2012, 2014), who drew vastly on the work of researchers from mainstream education and cognitive psychology (Baddeley, Eysenck, & Anderson, 2009; Dehn, 2008; Gagné et al., 1993). In his model, he illustrated how L2 input from the assessment environment is first processed in the short-term memory (STM), how the working memory (WM) retrieves and activates different types of knowledge from the long-term memory (LTM), and how all types of information are organized to produce a response. All of those steps appear to be controlled by the regulatory processes as seen in Figure 1.

Building upon his model of the architecture of human information processing, Purpura (2012) also explained how information processing might be applied to L2 processing when examinees are taking L2 assessments, and how examinees might resort to strategy use during each stage of information processing. As Figure 2 illustrates, the test input (information from the assessment in forms of texts, questions or prompts) is noticed and understood by employing comprehending processes (e.g., attending, decoding). The new input information is then encoded and temporarily held in WM by utilizing storing/remembering processes. This activates L2 knowledge and other types of knowledge in LTM, which is accessed and retrieved by means of retrieval processes, and temporarily stored again in WM until a response can be organized and eventually generated by utilizing output processes. After hearing the response, feedback from both self and others is generated so that improvement to the response can be made. While the test takers are engaged in all the stages of L2 information processing (i.e., input processing, central processing and output processing) to generate responses, a range of strategies are potentially involved throughout each stage. The strategies can be distinguished into four types according to Oxford's (2011) taxonomy of language learning strategies (as cited in Purpura, 2014, p. 21). The meta-cognitive, meta-affective, meta-sociocultural and meta-interactive strategies are respectively designed to regulate cognitive (e.g., revising), affective (e.g., coping), socio-cultural (e.g., cooperating) and interactive ones (e.g., managing turn-taking).

FIGURE 1
The architecture of human information processing. Adapted from Purpura (2012)

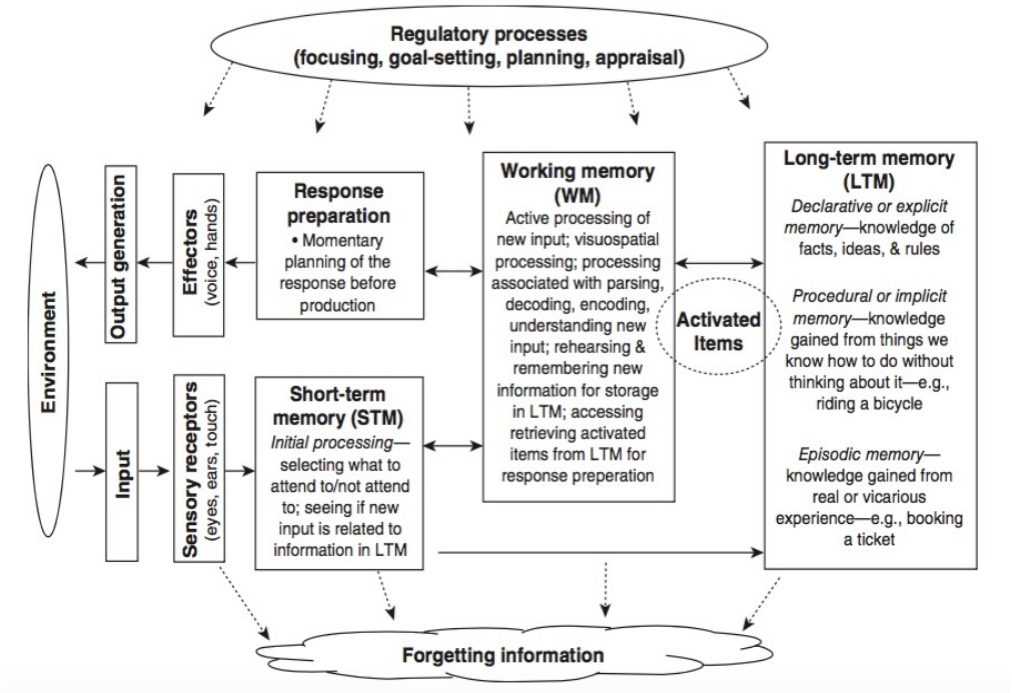
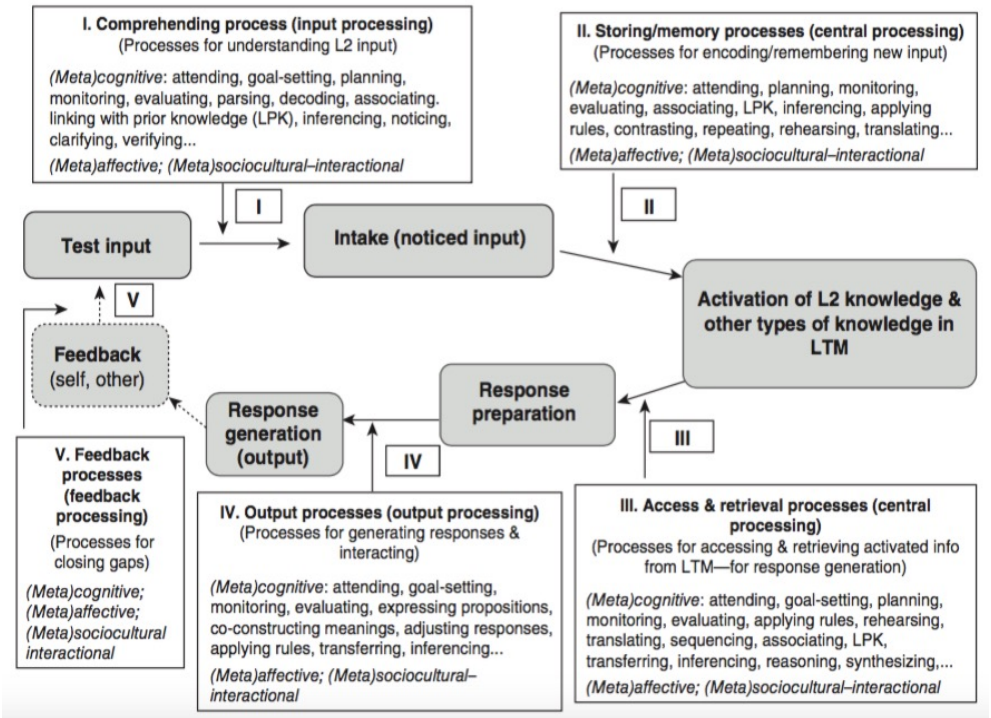


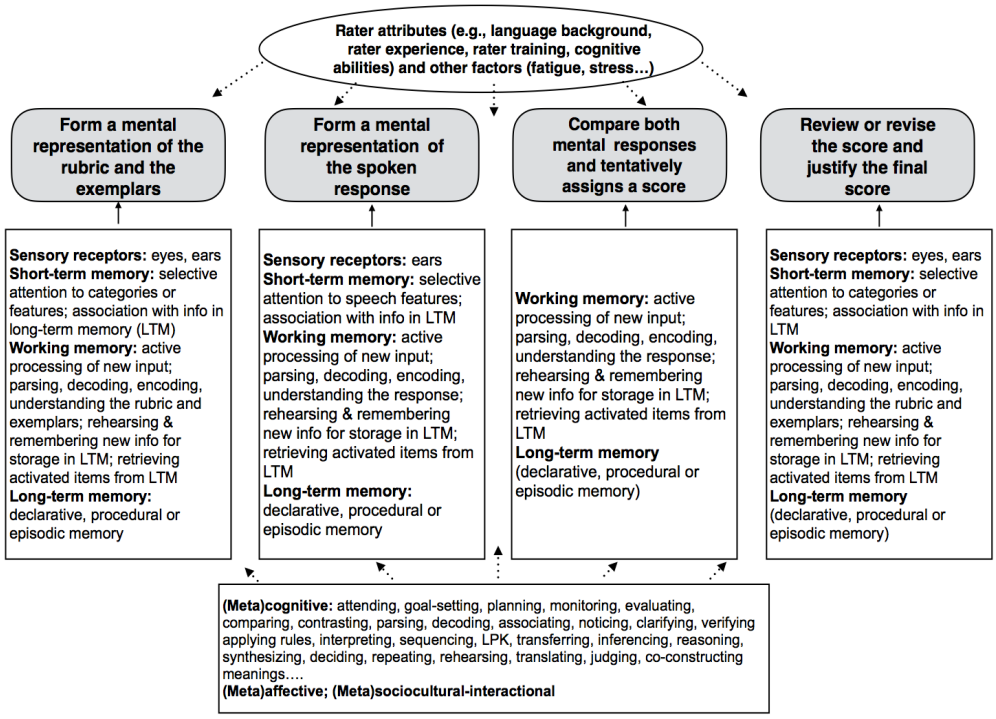
FIGURE 2
The interface of cognitive competence and L2 processing in assessment. Adapted from Purpura (2012)



Toward a cognitive model of rater cognition in L2 speaking assessment

Drawing on Purpura's (2012) models of the architecture of human information processing and the interface of cognitive competence and L2 processing, as well as Bejar's (2012) model of the process of rating constructed responses, a tentative, unified model for hypothesizing rater cognition in the context of L2 speaking assessment is proposed here and will be discussed in terms of how it can assist future rater cognition research. As is shown in Figure 3, the model interfaces the process of rating L2 speech with the components of the architecture of human information processing that are activated and the cognitive processes that are invoked during each stage of rating. It also incorporates the wide range of (meta)cognitive strategies that raters may deploy to regulate the operations of their cognitive mechanisms during the rating process. This model delineates how the assessment input (i.e., the rubric, the exemplars, and the L2 spoken responses) might be picked up by sensory receptors and selectively attended to and initially processed in STM, how the WM decodes and encodes the assessment input information, retrieves and activates different types of knowledge from LTM, so that all the types of information can be reorganized and mental representations of both the rubric and the L2 responses can be formed. Then it describes how those mental representations are compared and contrasted to produce tentative scores in WM, and how the scores are reviewed or revised and justified through an iterative scoring process by utilizing all the cognitive components (i.e., sensory memory, STM, WM and LTM). All of those steps are regulated by the range of (meta)cognitive strategies (e.g., attending, monitoring) that are invoked and subject to the influences of various rater attributes.

FIGURE 3
A tentative, hypothesized cognitive-processing model of rating L2 spoken responses in assessment



The primary function of this tentative, hypothesized model is to postulate the nature of rater cognition which is underpinned by theories about the architecture of human information process and strategy use research in relation to external contexts. It is important to note that variability or differences in information processing within and among individuals can be affected not only by 1) individual rater characteristics, such as rating experience, training, age, gender, native language, cultural background, attitude toward L2 accents, and other personality variables such as cognitive and rating styles, but also 2) environmental variables, such as the characteristics of tasks, settings and environments in which rating or processing occurs.

The model is conceptualized from a cognitive-processing perspective and draws on some of the most up-to-date theories and frameworks on the human rating process and the architecture of human information processing in L2 assessment. It first serves as a summary of the essential findings from L2 speaking assessment that reflect what goes on in raters' minds during scoring. For instance, the various studies which have examined raters' attention to both criterion and non-criterion aspects and features of spoken performance (e.g., Ang-Aw & Goh, 2011; Brown, Iwashita, & McNamara, 2005; May, 2006, 2011; Orr, 2002) have revealed how raters have conceptualized the construct of oral proficiency before rating, which may have served as the basis for them to developmental representations of both the scoring rubric and examinees' responses during rating. The different scoring approaches that raters were found to adopt (e.g., Ang-Aw & Goh, 2011; Brown, 2000; May, 2006; Politt & Murray, 1996) also describe in broad terms how raters may have compared both mental representations in their mind and decided on scores about the performance. The influences of rater attributes (e.g., rater language background and rater experience) have also been suggested by the various studies reviewed on the effects of rater background variables (e.g., Zhang & Elder, 2011, 2014; Winke, Gass, & Myford, 2011, 2012; Wei & Llosa, 2015; Kim, 2011, 2015; Davis, 2012, 2015; Isaacs & Thompson, 2013) on their rating process. Instances of utilizing certain (meta)cognitive strategies (e.g., inference, comparing) were also reported in some studies (e.g., Brown, 2000; May, 2006; Orr, 2002). Brown (2000), for example, has found that raters frequently made inferences regarding certain scoring criteria and candidate characteristics. These inferences often differ from rater to rater, and are typically used to explain certain patterns of behavior or justify certain scores and decisions. Another example is that raters were found to make comparisons between examinees which might have influenced their judgments and decisions on scores (May, 2006; Orr, 2002).

Evaluating this model with empirical data is also a way forward. Since the model draws heavily on existing theories and research findings on the architecture of human information processing and (meta)cognitive strategy use in cognitive psychology (e.g., Baddeley, 2012; Baddeley, Eysenck, and Anderson, 2009), mainstream education (e.g., Dehn, 2008; Gagné, Yekovich, & Yekovich, 1993) and language assessment (e.g., Purpura, 2014), it is important to evaluate it through future empirical studies on the rating of L2 speaking performance, which allow for examination in greater detail the functions of different cognitive processes and (meta)cognitive strategies that appear to be involved in the process of rating. Research findings can be used to inform rater trainers about how raters' cognitive processes and strategy uses can be interpreted in light of the human information processing system, and how appropriate (meta)cognitive strategy use can contribute to improved rating performance.

To summarize, when raters are involved in the scoring process, the opportunity exists for multiple factors to influence the scores they produce in ways that present a threat to the validity of both human and automated scores. Research on rater cognition was reviewed to identify some of those threats. An attempt was made to organize aspects of rater cognition into a coherent

picture based on the research in L2 speaking performance assessment that is available, but also taking into account representative frameworks (e.g. Baddeley, 2012; Bejar, 2012; Purpura, 2012) in L2 assessment, main-stream education and cognitive psychology. A tentative, unified model of the process of scoring L2 spoken responses was proposed with an added call for further empirical verification through in-depth research adopting a cognitive-processing approach on rater cognition.

REFERENCES

Andringa, S. J., Hulstijn, J. H., Schoonen, R., van Beuningen, C. G., & Olsthoorn, N. M. (2012). *Determinants of successful listening proficiency*. Paper presented at the American Association for Applied Linguistics (AAAL), Boston.

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. *RELC journal*, *42*(1), 31-51.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3), Available from http://www.jtla.org.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychological Press.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology*, *63*, 1-29.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

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

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

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

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports*, *3*, 49-84.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, *2005*(1), i-157.

Cohen, A. D., & Upton, T. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, *24*(2): 209–50.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* (Unpublished doctoral dissertation). University of Hawai'i at Manoa, Hawaii.

Davis, L. (2015). The influence of training and experience on rater performance in scoring

spoken language. *Language Testing*, *33*(1), 117-135.

Dehn, M. J. (2008). *Working memory and academic learning*. Hoboken, NJ: John Wiley & Sons.

Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 215–242). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition*. Frankfurt: Peter Lang.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423-443.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221.

Erdosy, M. U. (2003). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *ETS Research Report Series*, *2003*(1), i-62.

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). New York: Longman.

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The Cognitive Psychology of School Learning*. New York, NY: Harper Collins College Publishers.

Greatorex, J., & Suto, W. I. (2006). *An empirical exploration of human judgement in the marking of school examinations*. Paper presented at the International Association for Educational Assessment Conference, Singapore.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly*, *9*, 186–203.

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 47-74.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second-language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*, 555–82.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*(1), 113.

Joe, J. N. (2008). *Using verbal reports to explore rater perceptual processes in scoring: An application to oral communication assessment*. ProQuest.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*(3), 239-258.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York, NY.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing*, *28*(2), 179–200.

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

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; How we score them. *ETS R & D Connections*, *11*.

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.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*, 54–71.

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing (MPLT)*, *11*(1), 29-51.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397-421.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–49). Cambridge, New York: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes* (Unpublished master's thesis). California State University Los Angeles, Los Angeles, California.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Studies in Language Testing*, *3*, 92-111.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1995). Reader calibration and its potential role in equating for the Test of Written English. *ETS Research Report Series*, *1995*(2), i-64.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system. *ETS Research Report Series*, *2000*(1), i-51.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

Oxford, R. L. (2011). *Teaching and researching language learning strategies*. London, England: Pearson.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*, 219–233.

Phakiti, A. (2003a). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26–56.

Phakiti, A. (2003b). A closer look at gender and strategy use in L2 reading. *Language Learning*, *53*(4), 649–702.

Phakiti, A. (2007). *Strategic competence and EFL reading test performance*. New York, NY: Peter Lang.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 74- 91). Cambridge, England: Cambridge University Press.

Purpura, J. E. (1997). An analysis of the relationships between test-takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, *47*(2), 289–325.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modeling approach. *Language Testing*, *15*(3), 333–79.

Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?* Paper presented at the American Association for Applied Linguistics Conference, Boston, MA.

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-25). John Wiley & Sons, Inc.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129-152). Cambridge University Press.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors* (Unpublished doctoral dissertation). Toronto: University of Toronto.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. *Studies in Immigrant English Language Assessment*, *1*, 159-189.

Stanovich, K. & West, R. (2002). Individual differences in reasoning. In T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Suto, I., Crisp, V., & Greatorex, J. (2008). Investigating the judgmental marking process: An overview of our recent research. *Research Matters: A Cambridge Assessment Publication, 5*, 6-8.

Suto, W. M. I., & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication, 2*, 7-11.

Suto, W. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*(2), 213-233.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential

analysis. *Journal of Educational Psychology*, *96*(1), 19–30.
Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283-304.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.
Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series*,(2), i-67.
Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.
Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 37-53.
Wolfe, E. W. (1995). A study of expertise in essay scoring (Unpublished doctoral dissertation). University of California, Berkeley.
Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83-106.
Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.
Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*, 37-56.
Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37.
Xi, X., & Mollaun, P. (2009). How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?. *ETS Research Report Serie*s, *2009*(2), i-37.
Xi, X., & Mollaun, P. (2011). Using Raters From India to Score a Large-Scale Speaking Test. *Language Learning*, *61*(4), 1222-1255.
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: competing or complementary constructs? *Language Testing*, *28*(1), 31-50.
Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgments of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325.