Journal of Contemporary Issues in Business and Government Vol. 27, No. 2,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Rating quality in rater mediated language assessment: a systematic literature review

Muhamad Firdaus MohdNoh[a], Mohd Ewan @ Effendi MohdMatore[a,*], NiusilaFaamanatu-Eteuati[b] and NorhidayuRosman[c]

[a]Faculty of Education, UniversitiKebangsaan Malaysia, Bangi, Malaysia
[b]Victoria University of Wellington, Wellington, New Zealand
[c]SekolahKebangsaanBugaya, Semporna, Malaysia

How has academic research in rating quality evolved over the last decade? Time has witnessed that previous researchers actively contributed to the development of knowledge particularly in ascertaining educational assessment to be updated with latest research-based practices. Thus, this review seeks to provide a bird's-eye view of research development on rating quality over the last ten years focusing on factors influencing raters' rating quality within the context of rater-mediated language assessment. This systematic literature review was conducted with the aid of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines through three stages, namely identification, screening and eligibility. Accordingly, the searching process has resulted in 43 articles to be thoroughly reviewed retrieved from two powerful database, Scopus and World of Science (WoS). Five major factors have emerged in response to the objective and they include rating experience, first language, rater training, familiarity and teaching experience. Analysis indicated that these factors lead to contradicting findings in terms of raters' rating quality except for rater training factor. Only rater training was proven to be successful in mitigating rater effect and enhancing raters' variability, severity and reliability. However, other factors were discovered to be inconclusive depending on whether they leave any impact on raters' rating quality. The direction for future studies is also discussed suggesting the inclusion of more qualitative or mixed-method studies conducted to be reviewed using other possible techniques.

Keywords: rating quality, rater-mediated assessment, rater variability, rating indicators, language assessment

1 Introduction

Rating quality is a fundamental issue in rater-mediated assessment. It is the central factor in determining the success of rating procedure because items used in the

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

assessment are rated subjectively asratersuse their professionalism and expertise to arrive at the final awarding marks (Eckes, 2019). Unlike objectively rated items, raters are not given list of acceptable answers rather a general guideline usually in the forms of rubric or rating scale with descriptors. Raters have the freedom to accept answers that match with the guidelines provided by test developers. Hence, candidates' observed marks are not merely a reflection of their abilities in the assessed domains but a combination of their abilities and raters' expertise (Engelhard & Wind, 2018). However, ratings provided by human raters are not free from irrelevant influence. Raters' idiosyncrasies brought into rating procedures can be impactful in positive or negative ways. The discussion of rating quality in rater-mediated assessment is contentious when it is executed for high-stakes setting because the results are used to make significant decision in candidates' lives. High-stakes assessment accentuates the principles of quality measurement including validity, reliability and fairness in the interpretation of candidates' results. Achievement of measurement principles in rater-mediated assessment is thus salient to ensure that candidates are only assessed based on their true abilities.

Systematic literature review (SLR) is a method of collecting and processing information from all accessible data to guide the researcher to answer research questions (Petticrew & Roberts, 2006). It is a technique used to identify what has been proposed by previous researchers, what is proven to be working or not in relation to an identified problem. Among the benefits of conducting a systematic literature review is the selection process of which articles to be reviewed will result in reviewing only significant articles amidst the information overload. Meanwhile, literature on rating quality has been reviewed by previous researchers with different focus. Reviewing articles on cognitive process of raters in assessing second language speaking assessment, Han (2016) has categorized the studies into how and why raters differed in their ratings. How raters generate different ratings was discussed through features that raters paid attention to, raters' approach of rating either holistically or analytically and raters' treatment of criteria and non-criteria relevant aspects. On the other hand, why raters differed in their ratings was discussed through the effects of raters'language background familiarity, rating experience and rater training. Researchers have generally concurred that these three types of background leave significant impact on raters' cognitive processes and rating behaviour in terms of how they gave comments, scoring features they focus on, interpretation of scoring domains and how they decide for final scores. Whereas, another systematic review by Wind & Peterson (2018) has been published to analyse the types of methods used in evaluating rating quality. In their methodological meta-analysis, they reviewed the kinds of statistical methods researchers employed and the types of indices of rating quality to report how sample raters performed in the studies.

While these two systematic reviews are contributing to the growing discussion of rating quality in rater-mediated assessment, the present study aims at systematically reviewing literature from 2010 to 2020. The ten-year time frame is important as to

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

inform researchers and practitioners particularly in educational assessment on the recent discovery obtained by previous researchers. In particular, the objective of this reviewis to identify significant factors influencing rating quality in the ratings produced by raters mostly researched. The central focus of this systematic literature review is on the practice of rating among raters in language assessment either in speaking or writing assessment because rater-mediated assessment is ubiquitous in language testing. In fact, most of the studies conducted to investigate rating quality are executed within the context of language assessment apart from other contexts such as musical and creativity assessment. The current study is salient in offering an updated discussion on rating quality by highlighting what factors (independent variables) lead to raters' rating quality in educational assessment. Additionally, the review is further supported from the robust process in SLR to warrant the analysis of only high-quality research studies with empirical data.

2 METHODOLOGY

The method used to review the selected articles was discussed including publication standards, database, eligibility and exclusion criteria, stagesin the review process (identification, screening, eligibility) as well as data abstraction and analysis.

2.1 Publication Standards

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) is employed as the publication standard in this review. PRISMA guidelines were established with the intention of improving and standardizing the quality of reporting in systematic reviews (Fleming, Koletsi, & Pandis, 2014). The guidelines can assist to enhance transparency of reporting and reviewing articles apart from help researchers in establishing valid and reliable findings and conclusions (McInnes et al., 2018). Particularly in this study, it offers guideline to conduct rigorous search of terms related to rating quality particularly in rater-mediated language assessment.

2.2 Database

Articles reviewed in this study were extracted from two powerful databases accessed online, namely Scopus and World of Science (WoS). Scopus gathers more than 30,000 journals from more than 11,000 publishers and are peer-reviewed covering top-level subject fields including education, measurement and language. Each journal is reviewed annually to maintain the standards and the quality that are evaluated based on their h-index, CiteScore, SCImago Journal Rank and Source Normalized Impact per Paper (SNIP). It means any Scopus article is of high quality and goes through a thorough process before being accepted to be published. Meanwhile, WoS is a website which offers access to robust database to plethora of academic disciplines. It is currently monitored by Clarivate Analytics but was originally created by the Institute for Scientific Information (ISI). It covers more than 30,000 journals with more than 250 disciplines.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

2.3 Systematic Literature Review Stages

This review was conducted through three main stages – identification, screening and eligibility. In the first stage, keywords to be employed in the searching process were identified. These keywords were related to the research questions in this review as well as those suggested by previous literature and thesaurus. The main keywords like rating quality, rater effects and rater-mediated assessment were used (see Table 1).

Table 1. The search string used for the systematic literature review process.

Database	Keywords
Scopus	TITLE-ABS-KEY (("rating quality" OR "rating accuracy" OR "rating performance" OR "rater effect*" OR "rater bias" OR "rater performance*" OR "rater behavior*" OR "rater accuracy" OR "rater cognition" OR "rater error" OR "rater difference*" OR "rater prejudice" OR "rater reliability" OR "rater consistency" OR "rater training" OR "rater background" OR "rater expert*" OR "rater characteristic*" OR "rater experience" OR "rater perception" OR "rater knowledge" OR "rater familiarity" OR "rater cognitive" OR "rater agreement" OR "rater competence" OR "rater evaluation") AND ("educational assessment" OR "language" OR "language assessment" OR "writing assessment" OR "writing test*" OR "speaking test*" OR "oral test*" OR "speaking assessment" OR "rater-mediated assessment"))
Web of Science	(TS=(("rating quality" OR "rating accuracy" OR "rating performance" OR "rater effect*" OR "rater bias" OR "rater performance*" OR "rater behavior*" OR "rater accuracy" OR "rater cognition" OR "rater error" OR "rater difference*" OR "rater prejudice" OR "rater reliability" OR "rater consistency" OR "rater training" OR "rater background" OR "rater expert*" OR "rater characteristic*" OR "rater experience" OR "rater perception" OR "rater knowledge" OR "rater familiarity" OR "rater cognitive" OR "rater agreement" OR "rater competence" OR "rater evaluation") AND ("educational assessment" OR "language" OR "language assessment" OR "writing assessment" OR "writing test*" OR "speaking test*" OR "oral test*" OR "speaking assessment" OR "rater-mediated assessment")))

These keywords are chosen because they were prominently used by previous researchers to classify their studies. The searching process was conducted using the two databases mentioned earlier. Then, in the second stage, the screening was carried out to choose only the significant articles to be reviewed. Several criteria were used as filter (see Table 2).

Journal of Contemporary Issues in Business and Government Vol. 27, No. 2,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Table 2. The filtering criteria

Criteria	Included	Excluded
Language	English	Language other than English language
Document types	Journal articles	Conceptual paper, systematic review, book chapter, proceedings, thesis, report
Time frame	2010 - 2020	< 2010
Subject area	Education, social science, humanities	Medicine, Mathematics, Health, Dentistry, Engineering
Article types	Research articles with empirical data	Methodological articles, systematic review, item development, peer-assessment, special education

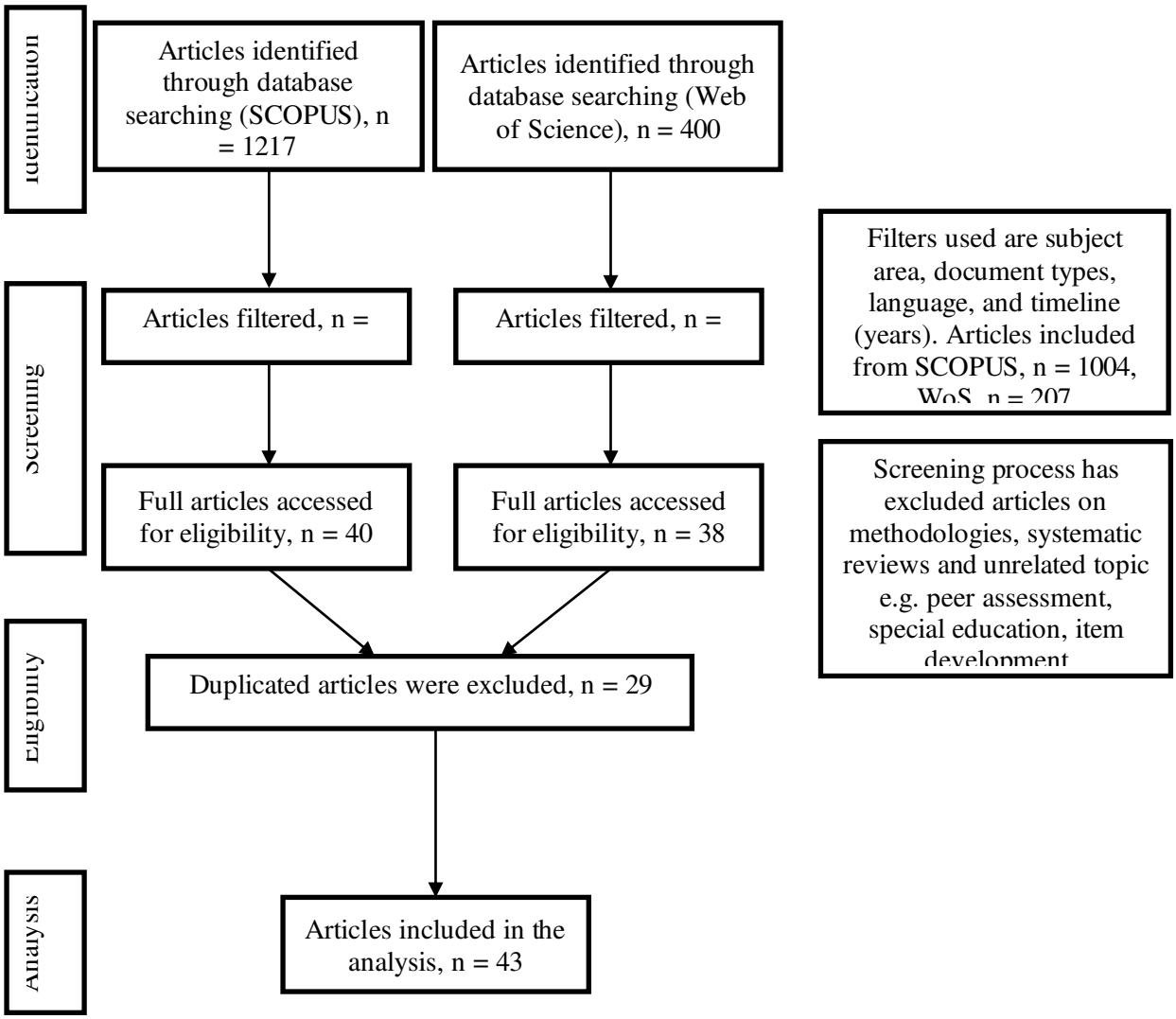

Fig. 1. Flow diagram of the study

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Firstly, articles in language other than English were filtered and excluded from the list so that articles can be reviewed without the need to translate them. Secondly, the type of publication selected were only articles with empirical data from academic journals because this type of paper reports studies conducted based on well-design methods by involving education stakeholders such as teachers and students. It means other publications such as report, theses, article review, and conceptual papers are excluded from the list. Thirdly, only articles from 2010 to 2020 were selected. This time frame was important as to observe how the development of knowledge and publication in rater-mediated assessment has progressed over the last ten years. Finally, since the focus of this systematic literature review was on language assessment, only articles published within the context of language assessment were chosen. Next, in the third stage, the eligibility criteria through which 65 articles from Scopus and 59 articles from WoSwere accessed. After a thorough scrutiny, only 43 articles were reviewed (see Fig. 1).

2.4 Data abstraction and analysis

The selected articles were then read through and analysed. The articles were reviewed first by perusing on the abstract. Later, the exploration of full articles to identify subthemes under the two themes identified in the research questions – factors influencing raters and indicators to manifest raters' rating quality. Content analysis was utilized to identify the subthemes related to each theme.

3 RESULTS

3.1 General findings

According to the years of publication, 2011 and 2016 have the greatest number of articles published with seven articles, followed by 2014 with six articles, four articles respectively in 2013, 2017, 2018 and 2019, three articles in 2015, two articles in 2010 and finally one article in 2012 and 2020 (as of March 2020) as shown on Fig. 2.

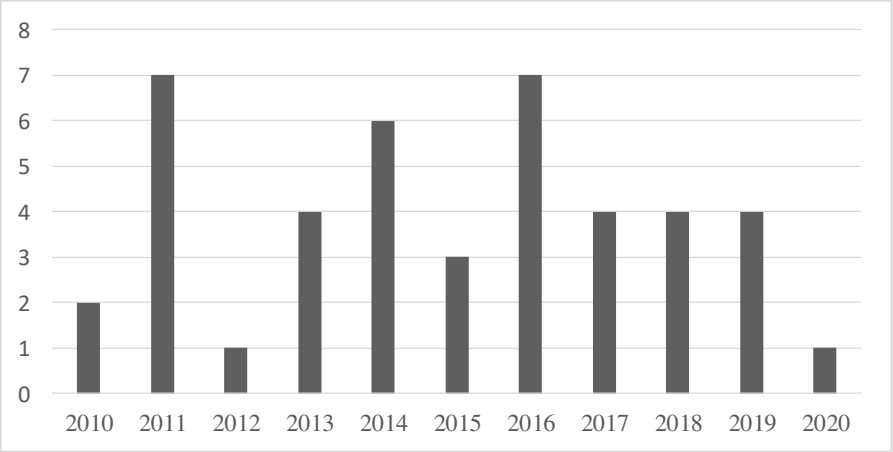

Fig. 2. *Number of articles based on publication year*

Journal of Contemporary Issues in Business and Government Vol. 27, No. 2,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

In terms of geographical location and where the studies have been conducted, 16 countries were identified. United States of America (USA) has the most number of articles published with 13 articles, followed by Iran (seven articles), Canada (five articles), Turkey and United Kingdom (three articles), Taiwan (two articles) and one article respectively in Netherlands, Korea, Sweden, Japan, Estonia, China, Brazil, Singapore, Australia and India as presented on Fig. 3.

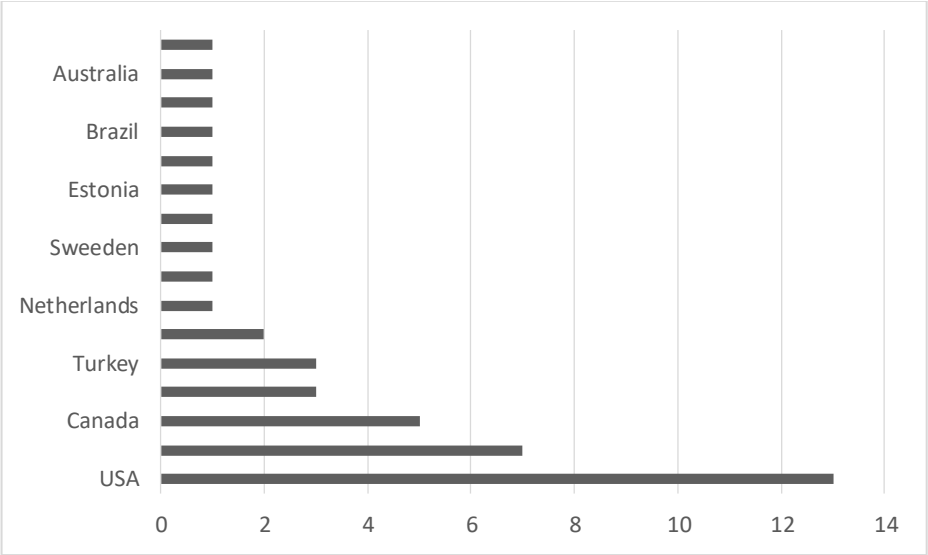

Fig. 3. *Number of articles based on countries*

Regarding the study design, 31 studies were conducted using quantitative design, 11 studies used mixed method design and only one study was carried out qualitatively. As for the context of language assessment, the studies were divided into two different contexts namely the speaking assessment (29 studies) and the writing assessment (14 studies).Five major factors have emerged to answer the first research objective. As presented on Table 3, the five factors are rating experience (13 articles), first language (12 articles), rater training (11 articles), familiarity with candidates (11 articles) and teaching experience (8 articles). It was indicated that most researchers employed quantitative design and carried out their study within speaking assessment.

Table 3. The findings of the analysis

Author	Year	Skills		Design			Factors				
		Speaking	Writing	Quantitative	Qualitative	Mixed	Rating Experience	First Language	Rater Familiarity	Rater Training	Teaching Experience
Şahan & Razı (2020) – Turkey	2020		*	*			*				

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Kang et al. (2019) – USA	2019	*		*			*	*	*	*
Ahmadi Shirazi (2019) – Iran	2019		*	*		*	*			
Wikse Barrow et al. (2019) - Iran	2019	*		*				*		
Bijani (2019) – Iran	2019	*		*					*	
Bijani (2018) – Iran	2018	*			*				*	
Huang et al. (2018) – Taiwan	2018	*		*		*			*	
Seker (2018) – Turkey	2018		*	*					*	
Eckstein & Univer (2018) - USA	2018		*	*						*
Bijani & Khabiri (2017) – Iran	2017	*		*					*	
Duijm et al. (2017) - Netherlands	2017	*		*					*	
Tanriverdi-Koksal & Ortactepe (2017) - Turkey	2017	*		*				*		
Kang & Veitch (2017) – USA	2017		*	*						*
Attali (2016) – USA	2016		*	*		*			*	
Huang et al. (2016) – USA	2016	*		*				*		
Davis (2016) – USA	2016	*		*		*			*	
Saito & Shintani (2016) - Canada	2016	*		*				*		
Lee (2016) – Korea	2016		*	*						*
Marefat & Heydari (2016) - Iran	2016		*	*			*			
Sandlund & Sundqvist (2016) - Sweden	2016	*		*				*		
Kim (2015) – USA	2015	*			*		*		*	*
Hijikata-Someya et al. (2015) - Japan	2015		*			*		*		

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Wei & Llosa (2015) – USA	2015	*			*		*				
Huang & Jun (2014) – USA	2014	*		*			*				
Stassenko et al. (2014) - Estonia	2014	*			*						
Tajeddin & Alemi (2014) – Iran	2014		*	*			*		*		
Wester & Mayo (2014) – UK	2014	*		*			*				
Zhang & Elder (2014) – China	2014	*			*		*				
Schmid & Hopp (2014) - Brazil	2014	*		*				*			
Huang (2013) – USA	2013	*		*				*		*	
He et al. (2013) – Taiwan	2013		*	*							
Isaacs & Thomson (2013) - Canada	2013	*			*	*					
Winke, Gass, & Myford (2013) - USA	2013	*		*				*			
Kang (2012) – USA	2012	*		*			*			*	
Ang-Aw & Goh (2011) - Singapore	2011	*			*	*					
Barkaoui (2011) – Canada	2011		*		*	*	*				
Carey, Mannell, & Dunn (2011) - Australia	2011	*		*				*			
Hsieh (2011) – USA	2011	*			*					*	
Leckie & Baird (2011) – UK	2011	*		*		*					
Lim (2011) – UK	2011		*	*		*					
Xi & Mollaun (2011) – India	2011	*		*			*	*			
Barkaoui (2010) – Canada	2010		*		*	*					
Barkaoui (2010b) – Canada	2010	*			*	*					

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

TOTAL	29	14	31	1	11	13	12	11	11	8

3.2 Rating Experience

A total of 13 articles have reported to study the influence of rating experience on the rating quality produced by sampleraters(Ahmadi Shirazi, 2019; Ang-Aw &Goh, 2011; Attali, 2016; Barkaoui, 2010b, 2010a; Davis, 2016; Huang et al., 2018; Isaacs & Thomson, 2013; Kim, 2015; Leckie & Baird, 2011; Lee, 2016; Lim, 2011; Şahan & Razı, 2020). The discussion in this section is based on three criteria, which are how researchers have defined 'rating experience', how sample raters are categorized based on their rating experience and the significance of rating experience on raters' rating quality. 'Rating experience' has mainly been defined by measuring whether the sample raters have experience in rating language assessment. Previous researchers have defined this variable through three different ways. Firstly, rating experience has been referred to sample raters' experience in rating any language assessment in general or any grading work that raters have done before the study took place (Attali, 2016; Isaacs & Thomson, 2013; Kim, 2015; Leckie & Baird, 2011; Şahan & Razı, 2020). Secondly, raterswere considered experienced if they have played role as a rater in any specific language assessment like writing assessment (Ahmadi Shirazi, 2019; Barkaoui, 2010a, 2010b) and speaking assessment (Huang et al., 2018). Finally, several studies have also identified raters' experience as an examiner in specific assessment in line with the context of the studies such as TOEFL iBT speaking test (Davis, 2016), writing section of the Michigan English Language Assessment Battery (MELAB) (Lim, 2011) and O'Level oral examination (Ang-Aw & Goh, 2011).

Researchers were different in how they classified the sample raters based on rating experience. Four studies have divided the raters into only two groups, namely novice raters and experienced raters(Attali, 2016; Barkaoui, 2010a, 2010b; Lim, 2011). Meanwhile, three studies have differentiated the sample raters by the length of years they were experienced in rating. Şahan & Razı, (2020) grouped the raters into three groups; first group was raters with three to four years' experience, second group was raters with five to six years' experience and third group was raters with more than six years' experience. Dividing the raters into two groups, two studies grouped raters with less than five years' experience in one group and raters with more than five years' experience in another group (Ahmadi Shirazi, 2019; Ang-Aw & Goh, 2011). Whereas, another two studies grouped raters based on binary basis whether they have experience or not (L. Huang et al., 2018; Isaacs & Thomson, 2013). Another study did not categorize the raters into their rating experience but compared rating quality among the ratings produced in four different rating sessions (Davis, 2016).

The findings of these studies have revealed contradicting pattern. Firstly, it was found that rating experience is a determinant of good rating quality. New raterswere found to show more variability in their ratings than experienced raters(Kim, 2015; Lim, 2011). Some of the new raters practiced high severity level, while other new raterswere

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

too lenient but managed to rate with the same quality as ratings produced by expert raters (Lim 2011). Apart from that, novice raters tended to show problematic ratings and were not consistent as compared to ratings from experienced raters'whowere more consistent and stable (Kim 2015). Nevertheless, there were also new and experienced raters who failed to manifest a uniform rating and did not generate a clear pattern (Ang-Aw & Goh, 2011). Both groups of raters exhibited different levels of severity and consistency but the pattern was not clear. Meanwhile, experience raters were also reported to use high severity due to their improved experiences in rating as well as their developed critical and analytical cognition (Barkaoui, 2010b). They also gave more negative and in-depth comments, focused more on language accuracy and tended to add criteria other than listed in the rubrics (Barkaoui, 2010a). Secondly, five other studies have suggested that rating quality were differentiated by raters' rating experience. All these studies have found that the differences of rating quality between raters of different rating experiences were not significant but experienced raters managed to achieve slightly higher inter-rater agreement (Isaacs & Thomson, 2013). The expert raters tended to include criteria not stated in the rubrics (Leckie & Baird, 2011), while new raters showed limited variability among them and failed to use all the rating scales provided to discriminate candidates (Attali, 2016). The low-experienced raters displayed different rating behaviour compared to medium- and high-experienced raters when assessing scripts of distinct qualities (Şahan & Razı, 2020) and both groups of raters did not manifest a uniform ratings using holistic scoring method (Ahmadi Shirazi, 2019).

3.3 First Language

Raters' first language often featured in studies of rating quality in rater-mediated assessment. A total of 12 studies have sought to investigate if this factor leaves impact on the kinds of ratings raters produce. First language was referred to raters' native language without any reference made to candidates' language (B. H. Huang & Jun, 2014; O. Kang, 2012; O. Kang, Rubin, & Kermad, 2019; Tajeddin & Alemi, 2014; Zhang & Elder, 2014). Other studies have conditioned raters' first language by considering they have the same native language with candidates such as Japanese language (Hijikata-Someya et al., 2015), Persian language in Iran (Ahmadi Shirazi, 2019; Marefat & Heydari, 2016), Indian language (Wei & Llosa, 2015; Xi & Mollaun, 2009) and German, Finnish and Mandarin (Wester & Mayo, 2014). In regard to grouping strategies of the sample raters, all of the studies have compared native speaking (NS) raters and non-native speaking (NNS) raters except for one study that employed native raters to assess non-native candidates (Tajeddin & Alemi, 2014).

In terms of raters' severity level, distinct findings have emerged. Some NNS raters were found to practice higher severity in overall scoring (Huang & Jun, 2014; O. Kang et al., 2019; Marefat & Heydari, 2016; Zhang & Elder, 2014a) but some NNS raterswere reported to manifest the same level of severity level as NS raters(Wei & Llosa, 2015). It was also discovered from other study that NS raters use differential

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

severity level according to examinees' first language (Wester & Mayo, 2014). NS raters were found to give high marks to non-native examinees and low marks to native examinees in assessing examinees' accentedness. However, analysis of severity practiced at different domains has resulted in different findings. NNS raters were more severe in assessing 'grammar' than 'organization' (Marefat & Heydari, 2016) but lenient in assessing 'flexibility and appropriacy' (Zhang & Elder, 2014).

Secondly, raters' inter-rater reliability was reported to be inconsistent. Raters from NS and NNS groups were found to be able to achieve the same inter-rater reliability (B. H. Huang & Jun, 2014; Zhang & Elder, 2014). However, inter-rater reliability was also observed to be low (Hijikata-Someya et al., 2015; Marefat & Heydari, 2016) and high for NNS raters(Xi & Mollaun, 2011). Finally, raters also showed different rating process. Raters who share the same first language with examinees were evident to be better at identifying language production that was affected by examinees' first language (Wei & Llosa, 2015). NNS raters regarded 'organization' domain as the most difficult domain to be assessed while NS raters perceived 'grammar' domain as the most difficult.

3.4 Raters' Familiarity

Familiarity with candidates has been investigated if raters' knowledge of specific accent related to raters' familiarity with candidates' first language accents (Carey et al., 2011; B. Huang et al., 2016; B. H. Huang, 2013; O. Kang et al., 2019; Saito & Shintani, 2016; Schmid & Hopp, 2014; Winke et al., 2013; Xi & Mollaun, 2011). However, the research findings into the effect of familiarity with candidates' accent has been inconsistent and contradictory. Three studies have reported essentially identical findings that raters' familiarity did not affect the quality of ratings that raters produced. Bilingual Indian raters who share the first language as the candidates managed to assess the candidates as reliable and valid as near-native speaking ratersXi & Mollaun (2011). Familiarity with candidates' Indian-accented English did not make Indian raters more severe or lenient in their assessment. This finding was then confirmed by Huang (2013) who concluded that raters' familiarity with candidates' accent did not leave statistically significant impact on their evaluation even though raters perceived that their evaluation was influenced by their knowledge and experience with the accent. A similar finding was also discovered by Huang et al. (2016) that even though raters' familiarity with a particular accent gained through heritage and formal education process has assisted raters in scoring speech features among candidates but did not contribute to rater bias. On the other hand, another five studies discovered that raters' familiarity with candidates' first language is a legitimate factor in determining their rating quality. Raters were found to be more lenient and tended to give high marks when they knew candidates' first language and when raters live in the same country as the candidates (Carey et al., 2011). Additionally, when raters have formally learned candidates' first language (Winke et al., 2013), have experience with second language learning, were exposed to English language varieties (Saito & Shintani, 2016) and when they have

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

frequent contact with the non-native candidates' language (O. Kang et al., 2019). The varying degree of raters' familiarity and exposure to foreign accents of the candidates were concluded to be responsible for the existence of variability in raters' rating behaviour(Schmid & Hopp, 2014).

Raters' familiarity was also researched in regard toraters' knowledge of candidates' academic progress (Sandlund & Sundqvist, 2016; Tanriverdi-Koksal & Ortactepe, 2017; Wikse Barrow et al., 2019) and interestingly similar findings were obtained. Students' teachers who were more familiar with candidates through everyday interaction were found to exhibit higher level of severity as compared to external raters who did not have preconception of candidates' learning progress (Sandlund & Sundqvist, 2016). In another study, comparison between before and after raters were given information on candidates' proficiency levels revealed that raters tended to change their marks in the second rating sessions (Tanriverdi-Koksal & Ortactepe, 2017). Positively, knowledge on candidates' academic progress have assisted raters in judging candidates but rendering raters to score students differently as compared to when they were not given the information. Apart from that, raters' familiarity with candidates' language development through their vocational work as teachers have made their scoring more valid and reliable (Wikse Barrow et al., 2019).

3.5 Rater Training

Ideally, rater training is conducted prior to rating procedure to enable raters to rate with desirable rating quality. Rater training is claimed to be potent in alleviating the effect of irrelevant variables on raters' rating quality. A total of eleven articles have revealed the influence of rater training. The studies were conducted using two study designs, which are comparison of rating quality produced by raters before and after a training by assigning raters to rate at least two times (Bijani, 2018, 2019; Davis, 2016; O. Kang, Rubin, & Kermad, 2019; H. J. Kim, 2015; Seker, 2018). Furthermore, comparison of rating quality produced by raters differentiated by their experience in attending rater training before the study was conducted (Ang-Aw & Goh, 2011; Attali, 2016; Duijm et al., 2017; L. Huang et al., 2018; Kim, 2015). Consistent findings have emerged from all the studies that rater training played a significant role in determining the quality of ratings that raters produced. It was observed that raters improved in their rating quality after they attended rater training. After training, their inter-rater agreement has increased (Davis, 2016; Seker, 2018; Tajeddin & Alemi, 2014) and their dispersion index has decreased (Tajeddin & Alemi, 2014) which means they managed to rate with uniformity. Their severity and leniency have also moved closer to the mean score which made their rating as a group to be within the acceptable range of severity level (O. Kang et al., 2019).

When raters are divided into different groups based on the amount of rater training that they have attended, rater training carried out during the study was found to be more impactful for novice and developing raters as compared to experienced raters(Kim, 2015). Apart from that, one unanticipated finding was that rater training was

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

more influential than raters' existing experience received from previous rater training they attended before the study was implemented (Attali, 2016). It was proven when the observed difference between raters with different amount of previous rater training attended was not significant. Conversely, Duijm et al. (2017) reported significant difference in rating quality between professional raterswho have undergone training and non-professional raters. Non-professional raters tended to award high marks to candidates especially in 'fluency' domain.

3.6 Teaching Experience

Another significant factor emerged from this review is teaching experience found in eight articles. This factor has been operationally used by previous researchers to denote different meanings. Three studies regarded teaching experience by measuring the length of years sample raters have been teaching English language (O. Kang et al., 2019; Kim, 2015; Lee, 2016). Another three studies defined teaching experience by raters' status as a teacher by comparing rating quality produced by teachers and non-teachers (Hsieh, 2011; B. H. Huang, 2013; O. Kang, 2012). The remaining two studies differentiated raters' performance by comparing their teaching subjects (Eckstein & Univer, 2018; H. S. Kang & Veitch, 2017).

Analysis of rating quality in these studies have resulted in contradictory findings. Three studies have reported significant difference in rating quality between the compared groups (H. S. Kang & Veitch, 2017; O. Kang, 2012; Lee, 2016). It was discovered that teachers of different area of teaching assessed students with different rating quality – the non-science teachers mark students more lenient and the childhood education teachers are the most stringent (H. S. Kang & Veitch, 2017). A comparison between new lecturers and experienced lecturers has revealed that new lecturers were more lenient in their scores (Lee, 2016). Additionally, a regression analysis of raters' background has suggested that teaching experience was a potent variable in determining rating quality (O. Kang, 2012). Meanwhile, three other studies have concluded that some differences among raters of distinct teaching experience were observed particularly in terms of the assessment features that raters paid more attention to (Eckstein & Univer, 2018; B. H. Huang, 2013; O. Kang et al., 2019). For example, second language writing teachers preferred to put high value on rhetorical, lexical and grammatical features of students' text but teachers who are teaching English as first language tent to spend more time on originality and criticality of students' work. However, the differences were very minimal and proven to be statistically insignificant. Notwithstanding the insignificant difference, teacher raters in comparison with non-teacher raters were found to be better at evaluating candidates analytically by discriminating candidates' abilities based on different linguistic features (Hsieh, 2011; Huang, 2013) and were less influenced by candidates' foreign accents (Huang, 2013).

4 DISCUSSION

Rating quality is a global concern in any educational assessment system especially in

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

the context of rater-mediated assessment. The review has attempted to analyse the existing literature on factors influencing raters' rating quality in rater mediated assessment within the context of language testing over the last ten years from 2010 to 2020. Rating quality is the main consideration in rater-mediated assessment because of the subjectivity in ratings provided by human raters(Engelhard & Wind, 2018). It is inevitable that raters bring their own identities to the rating scenes which may negatively affect the quality of ratings they generate and eventually become a source of irrelevant disturbance to candidates' holistic scores. A rigorous analysis from two gigantic database of academic research has resulted in 43 articles critical to the research objective. The analysis has shown that many factors can affect the quality of ratings that the raters produce. Within the scope of the review, five factors investigated by previous researchers have emerged. However, inconsistent findings were discovered for all the factors.

Raters' rating experience include their past involvement as an examiner or assessor during the process of marking candidates' answers in a specific context of assessment type or general task of assessing students' work. Discussion on whether this factor is legitimate in determining raters'rating quality is inconclusive. Raters' rating experience is claimed to be supporting their capability to assess candidates with acceptable range of rating quality (Barkaoui, 2010b; Kim, 2015) especially in mitigating irrelevant factors to influence their ratings. Due to their varied experiences in rating, experienced raters are found to be rating with good severity, variability and consistency level (Lim, 2011). Rating experience may have developed their confidence in using the rating rubrics which enabled them to provide more valid ratings. In contrast, novice raters who were found to manifest varied severity and consistency level (Barkaoui, 2010a, 2010b) depended too much on rating rubrics even though they probably were less certain on how the rubrics work. Consequently, they were not able to provide uniform ratings among them. On the other hand, some studies have also reported that raters of different rating experience were able to rate with the same quality (Ahmadi Shirazi, 2019; Attali, 2016; Isaacs & Thomson, 2013; Şahan & Razı, 2020). This finding may be attributable to how "rating experience" has been operationalized and different contexts of assessment used in the studies. Indeed, different types of language tests may create different assessment setting which eventually affect the way raters scorecandidates' answers.

Raters' language background including their first language and their exposure to candidates' first language were also indecisive as a determinant of rating quality. It was suggested that raters with different language background rate with different level of severity (Huang & Jun, 2014; O. Kang et al., 2019; Marefat & Heydari, 2016; Zhang & Elder, 2014a) and inter-rater reliability low (Hijikata-Someya et al., 2015; Marefat & Heydari, 2016; Xi & Mollaun, 2011). The findings may also be due to raters with different language background having different expectations and perceptionsof the candidates. They were also reported to put different emphasis on different rating criteria. Native speaking raters focus more on organization but non-native speaking

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

raterspaid more attention on grammar (Marefat & Heydari, 2016). Apart from that, non-native speaking raters were found to be facing more difficulties when they were assigned to mark using holistic scoring as compared to analytical method (Hijikata-Someya et al., 2015). It means that scoring methods have become a moderating factor to the difference in raters' rating quality between native speaking raters and non-native speaking raters. This may suggest that analytical scoring method may have assisted non-native speaking raters in discriminating candidates based on identified criteria. At the same time, this is also in contrast to holistic method that requires raters to assess candidates in general. Nevertheless, some studies have also reported that raters with different language background managed to exhibit the same severity (Wei & Llosa, 2015) and inter-rater reliability (B. H. Huang & Jun, 2014; Zhang & Elder, 2014) levels. This indicates thatraters' comprehension of candidates' speech production is not restricted by raters' language backgroundespecially in oral testing. Even though they may differ in terms of language background, raters have attended sufficient training that enable them to rate with the expected quality.

Majority of previous research on the effect of raters' familiarity have concluded that this factor can significantly affect rating quality. Raters who are familiar with candidates were reported to be more lenient (Carey et al., 2011; O. Kang et al., 2019; Saito & Shintani, 2016; Sandlund & Sundqvist, 2016; Schmid & Hopp, 2014; Tanriverdi-Koksal & Ortactepe, 2017; Wikse Barrow et al., 2019; Winke et al., 2013). This may suggest that raters' knowledge of candidates' language and academic background can be a source of irrelevant construct variance in raters' rating quality. Anonymous rating may be the best solution to solve this problem by assigning raters who are not familiar with candidates' background.

Rater training is undoubtedly proven to be impactful in alleviating the irrelevant construct variance among raters. All the articles reviewed concluded that rater training was effective to enable raters to score candidates within the acceptable ranges of severity, variability and reliability. Indeed, the aims of rater training is to familiarize raters with rating procedures, learn how to interpret rating rubrics, be able to apply them based on candidates' answers and differentiate candidates based on rating scales. It was also discovered that rater training during the study was more significant than their previous exposure of rating procedure (Attali, 2016).

Research on teaching experience factor has resulted in contradicting findings. Significant difference in rating quality was found between raters of distinct teaching experience (Kang & Veitch, 2017; Kang, 2012; Lee, 2016) but other studies have also found out that raters' rating quality was not differentiated by their teaching experience (Eckstein & Univer, 2018; B. H. Huang, 2013; O. Kang et al., 2019. This contradiction may be attributable to the different operationalization of "teaching experience" and different assessment system employed in those research.

5 Conclusion

This systematic literature review has analysed the selected articles to fulfil the research

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

objective. Five major findings have emerged in relation to factors influencing raterson how they rate candidates particularly within the context of language assessment. Over the last decade, those factors have been actively investigated by previous researchers, which are rating experience, raters' first language, raters' familiarity with candidates, rater training and raters' teaching experience. Findings from the studies suggested that all the factors lead to different effect in terms of raters' rating quality except rater training that is proven to enhance rating quality. Research studies of other factors have resulted in contradicting findings based on how the researchers operationally define the factor, study designs, assessment items used and context of the language assessment. The findings offer valuable lessons especially for educational assessment practitioners such as teachers and assessment developers. The appointment of teachers as raters in assessing students in any type of assessment need to consider teachers' varied background. Apart from that, rater training should be carefully designed to expose raters with the potential threat of failure to maintain objectivity and reliability throughout the rating procedures. Analysis has also shown that majority of the studies were carried out using quantitative approach apart from mixed method and only one qualitative study. Future research should attempt to investigate raters' rating quality through qualitative and mixed-method design as it offers an in-depth exploration and detailed discovery especially on what leads to raters' rating quality. Also, another rigorous review should consider incorporating other techniques such as citation tracking, reference crossing, snowballing, and contacting experts.

Acknowledgment

This work was supported by the Ministry of Higher Education (MOHE), Malaysia, and Faculty of Education, UniversitiKebangsaan Malaysia (UKM) through the Fundamental Research Grant Scheme (FRGS) under (Grant number: FRGS/1/2018/SSI09/UKM/02/1), and in part of Dana Penyelidikan FPEND (Grant number: GG-2019-034).The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped them improve the content, quality, and presentation of this article.

References

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open*, *9*(1), 1–14. https://doi.org/10.1177/2158244018822377

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, *42*(1), 31–51. https://doi.org/10.1177/0033688210390226

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

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, *33*(1), 1–20. https://doi.org/10.1080/2331186X.2018.1460901

Bijani, H. (2019). Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis. *Cogent Education*, *6*(1). https://doi.org/10.1080/2331186X.2019.1670592

Bijani, H., & Khabiri, M. (2017). Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment: A FACETS analysis. *Journal of Asia TEFL*, *14*(4), 687–702. https://doi.org/10.18823/asiatefl.2017.14.4.7.687

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2017). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501–527. https://doi.org/10.1177/0265532217712553

Eckstein, G., & Univer, B. Y. (2018). Assessment of L2 student writting : Does teacher disciplinary background matter? *Journal of Writing Research*, *10*(1), 1–23.

Engelhard, G., & Wind, S. A. (2018). *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. *Routledge*. New York & London: Routledge. https://doi.org/10.1017/CBO9781107415324.004

Fleming, P. S., Koletsi, D., & Pandis, N. (2014). Blinded by PRISMA: Are systematic reviewers focusing on PRISMA and ignoring other guidelines? *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0096407

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(1), 1–24. https://doi.org/10.7916/D8MS45DH

He, T.-H., Gou, W. J., Chien, Y.-C., Chen, I.-S. J., & Chang, S.-M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

assessment. *Psychological Reports*, *112*(2), 469–485. https://doi.org/10.2466/03.11.PR0.112.2.469-485

Hijikata-Someya, Y., Ono, M., & Yamanishi, H. (2015). Evaluation by native and non-native English teacher-raters of Japanese students' summaries. *English Language Teaching*, *8*(7), 1–12. https://doi.org/10.5539/elt.v8n7p1

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. Retrieved from https://michiganassessment.org/wp-content/uploads/2014/12/Spaan_V9_FULL.pdf#page=55

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41. https://doi.org/10.1080/15434303.2015.1134540

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*(3), 770–785. https://doi.org/10.1016/j.system.2013.07.009

Huang, B. H., & Jun, S. A. (2014). Age matters and so may raters: Rater differences in the assessment of foreign accents. *Studies in Second Language Acquisition*, *37*(4), 623–650. https://doi.org/10.1017/S0272263114000576

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, *8*(1), 1–17. https://doi.org/http://dx.doi.org/10.1186/s40468-018-0069-0

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. https://doi.org/10.1080/15434303.2013.769545

Kang, H. S., & Veitch, H. (2017). Mainstream teacher candidates' perspectives on ESL writing: The effects of writer identity and rater background. *TESOL Quarterly*, *51*(2), 249–274. https://doi.org/10.1002/tesq.289

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*(3), 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational*

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Measurement, *48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lee, K. R. (2016). Diversity among NEST raters: How do new and experienced NESTs evaluate Korean English learners' essays? *Asia-Pacific Education Researcher*, *25*(4), 549–558. https://doi.org/10.1007/s40299-016-0281-6

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

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, *27*(2016), 24–36. https://doi.org/10.1016/j.asw.2015.10.001

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., Clifford, T., … Willis, B. H. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies The PRISMA-DTA statement. *JAMA - Journal of the American Medical Association*, *319*(4), 388–396. https://doi.org/10.1001/jama.2017.19163

Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences*. *Systematic Reviews in the Social Sciences*. Malden: Blackwell Publishing. https://doi.org/10.1002/9780470754887

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*. https://doi.org/10.1177/0265532219900228

Saito, K., & Shintani, N. (2016). Foreign accentedness revisited: Canadian and Singaporean raters' perception of Japanese-accented English. *Language Awareness*, *25*(4), 305–317. https://doi.org/10.1080/09658416.2016.1229784

Sandlund, E., & Sundqvist, P. (2016). Equity in L2 English oral assessment: Criterion-based facts or works of fiction? *NJES Nordic Journal of English Studies*, *15*(2), 113–131. https://doi.org/10.35360/njes.365

Schmid, M. S., & Hopp, H. (2014). Comparing foreign accent in L1 attrition and L2 acquisition: Range and rater effects. *Language Testing*, *31*(3), 367–388. https://doi.org/10.1177/0265532214526175

Seker, M. (2018). Intervention in teachers' differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Educational Evaluation*, *59*(December 2017), 209–217. https://doi.org/10.1016/j.stueduc.2018.08.003

Stassenko, I., Skopinskaja, L., & Liiv, S. (2014). Investigating cultural variability in rater judgements of oral proficiency interviews. *Eesti Rakenduslingvistika Uhingu Aastaraamat*, (10), 269–281. https://doi.org/10.5128/ERYa10.17

Tajeddin, Z., & Alemi, M. (2014). Criteria and Bias in Native English Teachers' Assessment of L2 Pragmatic Appropriacy: Content and FACETS Analyses. *Asia-Pacific Education Researcher*, *23*(3), 425–434. https://doi.org/10.1007/s40299-013-0118-5

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.606

Tanriverdi-Koksal, F., & Ortactepe, D. (2017). Raters knowledge of students proficiency levels as a source of measurement error in oral assessments. *Hacettepe University Journal of Education*, *32*(3), 1–19. https://doi.org/10.16986/HUJE.2017027583

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. https://doi.org/10.1080/15434303.2015.1037446

Wester, M., & Mayo, C. (2014). Accent Rating by Native and Non-native Listeners. *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 7749–7753.

Wikse Barrow, C., Nilsson Björkenstam, K., & Strömbergsson, S. (2019). Subjective ratings of age-of-acquisition: Exploring issues of validity and rater reliability. *Journal of Child Language*, *46*(2), 199–213. https://doi.org/10.1017/S0305000918000363

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

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 Series*, *2009*(2), i–37. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306–325. https://doi.org/10.1080/0969594X.2013.845547