Language Testing in Asia

RESEARCH Open Access

Examining consistency among different rubrics for assessing writing

Enayat A. Shabani* and Jaleh Panahi

* Correspondence: eshabani@tums.ac.ir
Department of Foreign Languages, TUMS International College, Tehran University of Medical Sciences (TUMS), Keshavarz Blvd., Tehran 1415913311, Iran

Abstract

The literature on using scoring rubrics in writing assessment denotes the significance of rubrics as practical and useful means to assess the quality of writing tasks. This study tries to investigate the agreement among rubrics endorsed and used for assessing the essay writing tasks by the internationally recognized tests of English language proficiency. To carry out this study, two hundred essays (task 2) from the academic IELTS test were randomly selected from about 800 essays from an official IELTS center, a representative of IDP Australia, which was taken between 2015 and 2016. The test takers were 19 to 42 years of age, 120 of them were female and 80 were males. Three raters were provided with four sets of rubrics used for scoring the essay writing task of tests developed by Educational Testing Service (ETS) and Cambridge English Language Assessment (i.e., Independent TOELF iBT, GRE, CPE, and CAE) to score the essays which had been previously scored officially by a certified IELTS examiner. The data analysis through correlation and factor analysis showed a general agreement among raters and scores; however, some deviant scorings were spotted by two of the raters. Follow-up interviews and a questionnaire survey revealed that the source of score deviations could be related to the raters' interests and (un)familiarity with certain exams and their corresponding rubrics. Specifically, the results indicated that despite the significance which can be attached to rubrics in writing assessment, raters themselves can exceed them in terms of impact on scores.

Keywords: Scoring rubrics, Essay writing, Tests of English language proficiency, Writing assessment

Introduction

Writing effectively is a very crucial part of advancement in academic contexts (Rosenfeld et al. 2004; Rosenfeld et al. 2001), and generally, it is a leading contributor to anyone's progress in the professional environment (Tardy and Matsuda 2009). It is an essential skill enabling individuals to have a remarkable role in today's communities (Cumming 2001; Dunsmuir and Clifford 2003). Capable and competent L2 writers demonstrate their idea in the written form, present and discuss their contentions, and defend their stances in different circumstances (Archibald 2004; Bridgeman and Carlson 1983; Brown and Abeywickrama 2010; Cumming 2001; Hinkel 2009; Hyland 2004). Writing correctly and impressively is vital as it ensures that ideas and beliefs are

expressed and transferred effectively. Being capable of writing well in the academic environment leads to better scores (Faigley et al. 1981; Graham et al. 2005; Harman 2013). It also helps those who require admission to different organizations of higher education (Lanteigne 2017) and provides them with better opportunities to get better job positions. Business communications, proceedings, legal agreements, and military agreements all have to be well written to transmit information in the most influential way (Canseco and Byrd 1989; Grabe and Kaplan 1996; Hyland 2004; Kroll and Kruchten 2003; Matsuda 2002). What should be taken into consideration is that even well until the mid-1980s, L2 writing in general, and academic L2 writing in particular, was hardly regarded as a major part of standard language tests desirable of being tested on its own right. Later, principally owing to the announced requirements of some universities, it meandered through its path to first being recognized as an option in these tests and then recently turning into an indispensable and integral part of them.

L2 writing is not the mere adequate use of grammar and vocabulary in composing a text, rather it is more about the content, organization and accurate use of language, and proper use of linguistic and textual parts of the language (Chenoweth and Hayes 2001; Cumming 2001; Holmes 2006; Hughes 2003; Sasaki 2000; Weissberg 2000; Wiseman 2012). Essay, as one of the official practices of writing, has become a major part of formal education in different countries. It is used by different universities and institutes in selecting qualified applicants, and the applicants' mastery and comprehension of L2 writing are evaluated by their performance in essay writing.

Essay, as one of the most formal types of writing, constitutes a setting in which clear explanations and arguments on a given topic are anticipated (Kane 2000; Muncie 2002; Richards and Schmidt 2002; Spurr 2005). The first steps in writing an essay are to gain a good grasp of the topic, apprehend the raised question and produce the response in an organized way, select the proper lexicon, and use the best structures (Brown and Abeywickrama 2010; Wyldeck 2008). To many, writing an essay is hampering, yet is a key to success. It makes students think critically about a topic, gather information, organize and develop an idea, and finally produce a fulfilling written text (Levin 2009; Mackenzie 2007; McLaren 2006; Wyldeck 2008).

L2 writing has had a great impact on the field of teaching and learning and is now viewed not only as an independent skill in the classroom but also as an integral aspect of the process of instruction, learning, and most freshly, assessment (Archibald 2001; Grabe and Kaplan 1996; MacDonald 1994; Nystrand et al. 1993; Raimes 1991). Now, it is not possible to think of a dependable test of English language proficiency without a section on essay writing, especially when academic and educational purposes are of concern. Educational Testing Service (ETS) and Cambridge English Language Assessment offer a particular section on essay writing for their tests of English language proficiency. The independent TOEFL iBT writing section, the objective of which is to gauge and assess learners' ability to logically and precisely express their opinions using their L2 requires the learners to write well at the sentence, paragraph, and essay level. It is written on a computer using a word processing program with rudimentary qualities which does not have a self-checker and a grammar or spelling checker. Generally, the essay should have an introduction, a body, and a conclusion. A standard essay usually has four paragraphs, five is possibly better, and six is too many (Biber et al. 2004; Cumming et al. 2000). TOEFL iBT is scored based on the candidates' performance on two

tasks in the writing section. Candidates should at least do one of the writing tasks. Scoring could be done either by human rater or automatically (the eRater). Using human judgment for assessing content and meaning along with automated scoring for evaluating linguistic features ensures the consistency and reliability of scores (Jamieson and Poonpon 2013; Kong et al. 2009; Weigle 2013).

The Graduate Record Examination (GRE) analytic writing consists of two different essay tasks, an "issue task" and an "argument task", the latter being the focus of the present study. Akin to TOELF iBT, the GRE is also written on a computer employing very basic features of a word processing program. Each essay has an introduction including some contextual and upbringing information about what is going to be analyzed, a body in which complex ideas should be articulated clearly and effectively using enough examples and relevant reasons for supporting the thesis statement. Finally, the claims and opinions have to be summed up coherently in the concluding part (Broer et al. 2005). The GRE is scored two times on a holistic scale, and usually, the average score is reported if the two scores are within one point; otherwise, a third reader steps in and examines the essay (Staff 2017; Zahler 2011).

IELTS essay writing (in both Academic and General Modules) involves developing a formal five-paragraph essay in 40 min. Similar to essays in other exams, it should include an introductory paragraph, two to three body paragraphs, and a concluding paragraph (Aish and Tomlinson 2012; Dixon 2015; Jakeman 2006; Loughead 2010; Stewart 2009). To score IELTS essay writing, the received scores for the (four) components of the rubric are averaged (Fleming et al. 2011).

The writing sections of the Cambridge Advanced Certificate in English (CAE) and the Cambridge English: Proficiency (CPE) exams have two parts. The first part is compulsory and candidates are asked to write in response to an input text including articles, leaflets, notices, and formal and/or informal letters. In the second part, the candidates must select one of the writing tasks that might be a letter, proposal, report, or a review (Brookhart and Haines 2009; Corry 1999; Duckworth et al. 2012; Evans 2005; Moore 2009). The essays should include an introduction, a body, and a conclusion (Spratt and Taylor 2000). Similar to IELTS essay writing, these exams are scored analytically. The scores are added up and then converted to a scale of 1 to 20 (Brookhart 1999; Harrison 2010).

Assessing L2 writing proficiency is a flourishing area, and the precise assessment of writing is a critical matter. Practically, learners are generally expected to produce a piece of text so that raters can evaluate the overall quality of their performance using a variety of different scoring systems including holistic and analytic scoring, which are the most common and acceptable ways of assessing essays (Anderson 2005; Brossell 1986; Brown and Abeywickrama 2010; Hamp-Lyons 1990, 1991; Kroll 1990). Today, the significance of L2 writing assessment is on an increase not only in language-related fields of studies but also arguably in all disciplines, and it is a very pressing concern in various educational and also vocational settings.

L2 writing assessment is the focal point of an effective teaching process of this complicated skill (Jones 2001). A diligent assessment of writing completes the way it is taught (White 1985). The challenging and thorny natures of assessment and writing skills impede the reliable assessment of an essay (Muenz et al. 1999) such that, to date, a plethora of research studies have been conducted to discern the validity and reliability

of writing assessment. Huot (1990) argues that writing assessment encounters difficulty because usually, there are more than two or three raters assessing essays, which may lead to uncertainty in writing assessment.

L2 writing assessment is generally prone to subjectivity and bias, and "the assessment of writing has always been threatened due to raters' biasedness" (Fahim and Bijani 2011, p. 1). Ample studies document that raters' assessment and judgments are biased (Kondo-Brown 2002; Schaefer 2008). They also suggested that in order to reduce the bias and subjectivity in assessing L2 writing, standard and well-described rating scales, viz rubrics, should be determined (Brown and Jaquith 2007; Diederich et al. 1961; Hamp-Lyons 2007; Jonsson and Svingby 2007; Aryadoust and Riazi 2016). Furthermore, there are some studies suggesting the tendency of many raters toward subjectivity in writing assessment (Eckes 2005; Lumley 2005; O'Neil and Lunz 1996; Saeidi et al. 2013; Schaefer 2008). In light of these considerations, it becomes of prominence to improve consistency among raters' evaluations of writing proficiency and to increase the reliability and validity of their judgments to avoid bias and subjectivity to produce a greater agreement between raters and ratings. The most notable move toward attaining this objective is using rubrics (Cumming 2001; Hamp-Lyons 1990; Hyland 2004; Raimes 1991; Weigle 2002). In layman's terms, rubrics ensure that all the raters evaluate a writing task by the same standards (Biggs and Tang 2007; Dunsmuir and Clifford 2003; Spurr 2005). To curtail the probable subjectivity and personal bias in assessing one's writing, there should be some determined and standard criteria for assessing different types of writing tasks (Condon 2013; Coombe et al. 2012; Shermis 2014; Weigle 2013).

Assessment rubrics (alternatively called instruments) should be reliable, valid, practical, fair, and constructive to learning and teaching (Anderson et al. 2011). Moskal and Leydens (2000) considered validity and reliability as the two significant factors when rubrics are used for assessing an individual's work. Although researchers may define validity and reliability in various ways (for instance, Archibald 2001; Brookhart 1999; Bachman and Palmer 1996; Coombe et al. 2012; Cumming 2001; Messick 1994; Moskal and Leydens 2000; Moss 1994; Rezaei and Lovorn 2010; Weigle 2002; White 1994; Wiggan 1994), they generally agree that validity in this area of investigation is the degree to which the criteria support the interpretations of what is going to be measured. Reliability, they generally settle, is the consistency of assessment scores regardless of time and place. Rubrics and any rating scales should be so developed to corroborate these two important factors and equip raters and scorers with an authoritative tool to assess writing tasks fairly. Arguably, "the purpose of the essay task, whether for diagnosis, development, or promotion, is significant in deciding which scale is chosen" (Brossell 1986, p. 2). As rubrics should be conceived and designed with the purpose of assessment of any given type of written task (Crusan 2015; Fulcher 2010; Knoch 2009; Malone and Montee 2014; Weigle 2002), the development and validation of rating scales are very challenging issues.

Writing rubrics can also help teachers gauge their own teaching (Coombe et al. 2012). Rubrics are generally perceived as very significant resources attainable for teachers enabling them to provide insightful feedback on L2 writing performance and assess learners' writing ability (Brown and Abeywickrama 2010; Knoch 2011; Shaw and Weir 2007; Weigle 2002). Similarly, but from another perspective, rubrics help learners to follow a clear route of progress and contribute to their own learning (Brown and

Abeywickrama 2010; Eckes 2012). Well-defined rubrics are constructive criteria, which help learners to understand what the desired performance is (Bachman and Palmer 1996; Fulcher and Davidson 2007; Weigle 2002). Employing rubrics in the realm of writing assessment helps learners understand raters' and teachers' expectations better, judge and revise their own work more successfully, promote self-assessment of their learning, and improve the quality of their writing task. Rubrics can be used as an effective tool enabling learners to focus on their efforts, produce works of higher quality, get better grades, find better jobs, and feel more concerned and confident about doing their assignment (Bachman and Palmer 2010; Cumming 2013; Kane 2006).

Rubrics are set to help scorers evaluate writers' performances and provide them with very clear descriptions about organization and coherence, structure and vocabulary, fluent expressions, ideas and opinions, among other things. They are also practical for the purpose of describing one's competence in logical sequencing of ideas in producing a paragraph, use of sufficient and proper grammar and vocabulary related to the topic (Kim 2011; Pollitt and Hutchinson 1987; Weigle 2002). Employing rubrics reduces the time required to assess a writing performance and, most importantly, well-defined rubrics clarify criteria in particular terms enabling scorers and raters to judge a work based on standard and unified yardsticks (Gustilo and Magno 2015; Kellogg et al. 2016; Klein and Boscolo 2016).

Selecting and designing an effective rating scale hinges upon the purpose of the test (Alderson et al. 1995; Attali et al. 2012; Becker 2011; East 2009). Although rubrics are crucial in essay evaluation, choosing the appropriate rating scale and forming criteria based on the purpose of assessment are as important (Bacha 2001; Coombe et al. 2012). It seems that a considerable part of scale developers prefers to adapt their scoring scales from a well-established existing one (Cumming 2001; Huot et al. 2009; Wiseman 2012). The relevant literature supports the idea of adapting rating scales used in large-scale tests for academic purposes (Bacha 2001; Leki et al. 2008). Yet, East (2009) warned about the adaptation of rating scales from similar tests, especially when they are to be used across languages.

Holistic and analytic scoring systems are now widely used to identify learners' writing proficiency levels for different purposes (Brown and Abeywickrama 2010; Charney 1984; Cohen 1994; Coombe et al. 2012; Cumming 2001; Hamp-Lyons 1990; Reid 1993; Weir 1990). Unlike the analytic scoring system, the holistic one takes the whole written text into consideration. This scoring system generally emphasizes what is done well and what is deficient (Brown and Hudson 2002; White 1985). The analytic scoring system (multi-trait rubrics), however, includes discrete components (Bacha 2001; Becker 2011; Brown and Abeywickrama 2010; Coombe et al. 2012; Hamp-Lyons 2007; Knoch 2009; Kuo 2007; Shaw and Weir 2007). To Weigle (2002), accuracy, cohesion, content, organization, register, and appropriacy of language conventions are the key components or traits of an analytic scoring system. One of the early analytic scoring rubrics for writing was employed in the ESL Composition by Jacobs et al. 1981, which included five components, namely language development, organization, vocabulary, language use, and mechanics).

Each scoring system has its own merits and limitations. One of the advantages of analytic scoring is its distinctive reliability in scoring (Brown et al. 2004; Zhang et al. 2008). Some researchers (e.g. Johnson et al. 2000; McMillan 2001; Ward and McCotter

2004) contend that analytic scoring provides the maximum opportunity for reliability between raters and ratings since raters can use one scoring criteria for different writing tasks at a time. Yet, Myford and Wolfe (2003) considered the halo effect as one of the major disadvantages of analytic rubrics. The most commonly recognized merit of holistic scoring is its feasibility as it requires less time. However, it does not encompass different criteria, affecting its validity in comparison to analytic scoring, as it entails the personal reflection of raters (Elder et al. 2007; Elder et al. 2005; Noonan and Sulsky 2001; Roch and O'Sullivan 2003). Cohen (1994) stated that the major demerit of the holistic scoring system is its relative weakness in providing enough diagnostic information about learners' writing.

Many research studies have been conducted to examine the effect of analytic and holistic scoring systems on writing performance. For instance, more than half a century ago, Diederich et al. (1961) carried out a study on the holistic scoring system in a large-scale testing context. Three-hundred essays were rated by 53 raters, and the results showed variation in ratings based on three criteria, namely ideas, organization, and language. About two score years later, Borman (1979) conducted a similar study on 800 written tasks and found that the variations can be attributed to ideas, organizations, and supporting details. Charney (1984) did a comparison study between analytic and holistic rubrics in assessing writing performance in terms of validity and found a holistic scoring system to be more valid. Bauer (1981) compared the cost-effectiveness of analytic and holistic rubrics in assessing essay tasks and found the time needed to train raters to be able to employ analytic rubrics was about two times more than the required time to train raters to use the holistic one. Moreover, the time needed to grade the essays using analytic rubrics was four times the time needed to grade essays using holistic rubrics. Some studies reported findings that corroborated that holistic scoring can be the preferred scoring system in large-scale testing context (Bell et al. 2009). Chi (2001) compared analytic and holistic rubrics in terms of their appropriacy, the agreement of the learners' scores, and the consistency of rater. The findings revealed that raters who used the holistic scoring system outperformed those employing analytic scoring in terms of inter-rater and intra-rater reliability. Thus, there is research to suggest the superiority of analytic rubrics in assessing writing performance in terms of reliability and accuracy in scoring (Birky 2012; Brown and Hudson 2002; Diab and Balaa 2011; Kondo-Brown 2002). It is, generally speaking, difficult to decide which one is the best, and the research findings so far can best be described as inconclusive.

Rubrics of internationally recognized tests used in assessing essays have many similar components, including organization and coherence, task achievement, range of vocabulary used, grammatical accuracy, and types of errors. The wording used, however, is usually different in different rubrics, for instance, "task achievement" that is used in the IELTS rubrics is represented as the "realization of tasks" in CPE and CAE, "content coverage" in GRE, and "task accomplishment" in TOEFL iBT. Similarly, it can be argued that the point of focus of the rubrics for different tests may not be the same. Punctuation, spelling, and target readers' satisfaction, for example, are explicitly emphasized in CAE and CPE while none of them are mentioned in GRE and TOEFL iBT. Instead, idiomaticity and exemplifications are listed in the TOEFL iBT rubrics, and using enough supporting ideas to address the topic and task is the focus of GRE rating scales (Brindley 1998; Hamp-Lyons and Kroll 1997; White 1984).

Broadly speaking, the rubrics employed in assessing L2 writing include the above-mentioned components but as mentioned previously, they are commonly expressed in different wordings. For example, the criteria used in IELTS Task 2 rating scale are task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. These criteria are the ones based on which candidates' work is assessed and scored. Each of these criteria has its own descriptors, which determine the performance expected to secure a certain score on that criterion. The summative outcome, along with the standards, determines if the candidate has attained the required qualification which is established based on the criteria. The summative outcome of IELTS Task 2 rating scale will be between 0 and 9. Similar components are used in other standard exams like CAE and CPE, their summative outcomes being determined from 1 to 5. Their criteria are used to assess content (relevance and completeness), language (vocabulary, grammar, punctuation, and spelling), organization (logic, coherence, variety of expressions and sentences, and proper use of linking words and phrases), and finally communicative achievement (register, tone, clarity, and interest). CAE and CPE have their particular descriptors which demonstrate the achievement of each learners' standard for each criterion (Betsis et al. 2012; Capel and Sharp 2013; Dass 2014; Obee 2005). Similar to the other rubrics, the GRE scoring scale has the main components like the other essay writing scales but in different wordings. In the GRE, the standards and summative outcomes are reported from 0–6, denoting fundamentally deficient, seriously flawed, limited, adequate, strong, and outstanding, respectively. Like the GRE, the TOEFL iBT is scored from 0–5. Akin to the GRE, Independent Writing Rubrics for the TOEFL iBT delineates the descriptors clearly and precisely (Erdosy 2004; Gass et al. 2011).

Abundant research studies have been carried out to show that idea and content, organization, cohesion and coherence, vocabulary and grammar, and language and mechanics are the main components of essay rubrics (Jacobs et al. 1981; Schoonen 2005). What has been considered a missing element in the analytic rating scale is the raters' knowledge of, and familiarity with, rubrics and their corresponding elements as one of the key yardsticks in measuring L2 writing ability (Arter et al. 1994; Sasaki and Hirose 1999; Weir 1990). Raters play a crucial role in assessing writing. There is research to allude to the impact of raters' judgments on L2 writing assessment (Connor-Linton 1995; Sasaki 2000; Schoonen 2005; Shi 2001).

The past few decades have witnessed an increasing growth in research on different scoring systems and raters' critical role in assessment. There are some recent studies discussing the importance of rubrics in L2 writing assessment (e.g. Deygers et al. 2018; Fleckenstein et al. 2018; Rupp et al. 2019; Trace et al. 2016; Wesolowski et al. 2017; Wind et al. 2018). They commonly consider rubrics as significant tools for measuring L2 learners' performances and suggest that rubrics enhance the reliability and validity of writing assessment. More importantly, they argue that employing rubrics can increase the consistency among raters.

Shi (2001) made comparisons between native and non-native, as well as between experienced and novice raters, and found that raters have their own criteria to assess an essay, virtually regardless of whether they are native or non-native and experienced or novice. Lumley (2002) and Schoonen (2005) conducted comparison studies between two groups of raters, one group trained expert raters provided with no standard rubrics,

the other group novice raters with no training who had standard rubrics. The trained raters with no rubrics outperformed the other group in terms of accuracy in assessing the essays, implying the importance of raters. Rezaei and Lovorn (2010) compared the use of rubrics between summative and formative assessment. They argued that using rubrics in summative assessment is predominant and that it overshadows the formative aspects of rubrics. Their results showed that rubrics can be more beneficial when used for formative assessment purposes.

Izadpanah et al. (2014) conducted a study drawing on Jacobs et al. (1981) to see if the rubrics of one exam can be the predictor of another one. Practically, they wanted to examine whether the same score would be obtained if a rubric for an IELTS exam was used for assessing CPE or any other standard test. Their findings revealed that the rubrics were comparable with each other in terms of their different components by which different standard essays are assessed. Bachman (2000) compared TOEFL PBT and CPE and found a very meaningful relationship between the scores gained from essay writing tests. He also concluded that scoring CPE was usually more difficult than PBT, and that under similar conditions, exams from UCLES/Cambridge Assessment (like CPE) received lower scores in comparison to the ones from ETS (like PBT). In Fleckenstein et al. (2019) experts from different countries linked upper secondary students' writing profiles elicited in a constructed response test (integrated and independent essays from the TOEFL iBT) to CEFR level. The Delphi technic was used to find out the intra- and inter-panelist consistency while scoring students' writing profiles. The findings showed that panelists are able to provide ratings consistent with the empirical item difficulties and the validity of the estimate of the cut scores.

Schoonen (2005) and Attali and Burstein (2005) compared the generalizability of writing scores to different essays using only one set of the rubric. They checked and analyzed three components of writing rubric, including content, language use, and organization and found that the obtained scores from different essays are similar. Wind (2020) conducted a study to illustrate and explore methods for evaluating the degree to which raters apply a common rating scale consistently in analytic writing assessments. The results indicated a lack of invariance in rating scale category functioning across domains for several raters. Becker (2011) also examined different rubrics used to measure writing performance. He investigated the three different types of rubrics, namely holistic, analytic, and primary-trait scoring systems, to find which one is more appropriate for assessing L2 writing. He studied the merits and demerits of the three rubrics and concluded that none of them had superiority over the others, making each legitimate for assessing a piece of writing depending on the purpose of writing, the time allocated for assessment, and the raters' expertise.

In a recent study, Ghaffar et al. (2020) examined the impact of rubrics and co-constructed rubrics on middle school students' writing skill performance. The findings of their study indicated that co-constructed rubrics as assessment tools help students to outperform in their writing due to their familiarity with these types of rubrics. In addition, there are researchers who are of the contention that the use of rubrics is inconclusive and can be controversial especially when they are just used for summative assessment purposes and that when rubrics are used for both summative and formative assessment, they are more advantageous (Andrade 2000; Broad 2003; Ene and

Kosobucki 2016; Inoue 2004; Panadero and Jonsson 2013; Schirmer and Bailey 2000; Wilson 2006, 2017).

What all of these studies indicated is that employing well-developed rubrics increase equality and fairness in writing assessment. It is also suggested that various factors could affect writing assessment, especially raters' expertise and time allocated to the rating (Bacha 2001; Ghalib and Hattami 2015; Knoch 2009, 2011; Lu and Zhang 2013; Melendy 2008; Nunn 2000; Nunn and Adamson 2007). The purpose of the present study is twofold. First, it attempts to investigate the consistency among different standard rubrics in writing assessment. Second, it tries to examine whether any of these rubrics could be used as a predictor of others and if they all tap the same underlying construct.

Methods

Samples

To meet the objectives of the study, 200 samples of Academic IELTS Task 2 (i.e., essay writing) were used. The samples were randomly selected from more than 800 essays written as part of academic IELTS tests taken between 2015 and 2016 at an official IELTS test center, a representative of IDP Australia. The essays were asked to be written based on different prompts. As an instruction to the IELTS writing Task 2, it is required that the test takers write at least 250 words, a condition that 21 samples did not meet. Test takers were 19 to 42 years of age, 120 of the females and 80 males.

Raters

One of the raters in this study was an (anonymous) official IELTS examiner who had scored the essays officially; the other raters were four experienced IELTS instructors from an English department of a nationally prominent language institute, three males and one female, between 26 and 39 years of age, with 5 to 12 years of English language teaching experience. These four raters were selected based on their qualifications, teaching credentials and certifications, and years of teaching experience, particularly in IELTS classes. All the four raters were M.A. holders in TEFL and had been teaching different writing courses at universities and language institutes and were familiar with different scoring systems and their relevant components. Each rater was invited to an individual briefing session with one of the researchers to ensure their familiarity with the rubrics of interest and discuss some practical considerations pertaining to this study. They were asked to read and score each essay four times, each time based on one of the four rubrics (TOELF iBT, GRE, CPE, and CAE). The raters completed the scorings in 12 weeks during which time they were instructed not to share ideas about the task (the costs of scorings were modestly met).

Instrumentation

Four sets of rubrics for different writing tests (i.e., Independent TOEFL iBT, GRE, CPE, & CAE) were taken from ETS and Cambridge English Language Assessment. The official IELTS scores of the 200 essays were collected from the IELTS center. The rubrics employed for assessing and evaluating the writing tasks of these five standard exams

were analytic rubrics with different scales, namely a nine-point scale for assessing IELTS Task 2, five-point scales for GRE and TOEFL iBT, and six-point scales for CAE and CPE writing tasks. They assess the main components of essay writing construct, including the range of vocabulary and grammar used in addressing the task, cohesion and organization, and range of using cohesive devices, which were presented in different wordings in these rubrics.

Another instrument was a questionnaire designed by the researchers, which included both open-ended and closed-ended questions (see Appendix). The aim was to determine the raters' attitudes toward their rating experience and their familiarity with each exam and its corresponding rubrics. The themes of questionnaire items were determined based on a review of the literature on the important issues and factors affecting raters' performances and attitudes (Brown and Abeywickrama 2010; Coombe et al. 2012; Fulcher and Davidson 2007; Weigle 2002). In addition, an interview was carried out with the four raters to find out about their interest in rating and also to investigate their familiarity of the exams and their conforming rating scales.

Procedure

To carry out the study, 200 essay samples were scored once by a certified IELTS examiner. The assigned scores together with the IELTS examiner's relevant comments were written next to each essay sample. Afterward, all essays were rated by the four other raters, who were kept uninformed of the official IELTS scores. They were provided with the rubrics of the four essay writing tests and were instructed to assess each essay with the four given rubrics. By so doing, in addition to the official IELTS scores, four other scores were given to each essay from each rater; that is to say, each essay received 16 scores plus the official IELTS score. Therefore, all in all, the researchers collected 17 scores for each essay. The researcher-made questionnaire was carried out, and then an interview was conducted whereby the 4 raters were asked about their interest in rating and also their awareness and concerns about each exam and their relevant rubrics.

Analysis

To do the analysis of the data, the SPSS program, version 22, was employed. Initially, the descriptive statistics of the data were computed, and intercorrelations among the 17 scores were calculated to see if any statistically significant association could be found among the rubrics. To have a better picture of the existing association among the scoring rubrics of the different exams, PCA as a variant of factor analysis was run to examine the extent the rubrics tap the same underlying construct.

Results

To address the first research question, intercorrelations were computed among the IELTS, CAE, CPE, TOEFL iBT, and GRE scores. To answer the second research question, factor analysis was run to examine the extent the standard essay writings in these five tests of English language proficiency tap the same underlying construct. In this section, the results of the intercorrelations and factor analyses computations are reported in detail.

Intercorrelations among ratings

To estimate the intercorrelations among test ratings and raters, first, alpha was calculated for these five sets of scores together (i.e., IELTS, CAE, CPE, TOEFL iBT, and GRE). To analyze the data, primarily, alpha was calculated for each rater separately to check the consistency among raters. Then, alpha was computed for all the raters together to find inter-reliability among the raters. The intercorrelations were afterward computed between each exam score and the IELTS scores to see which score is (more) correlated with the IELTS.

Table 1 presents the alphas as the average of intercorrelations among the five sets of scores including the IELTS scores, and the four scores given by the raters. Evidently, rater 1 has an alpha of about .67, which is lower than the other alphas. However, because there were only five sets of scores correlated in each alpha, this low value of alpha could still be considered acceptable. Nevertheless, this lower value of alpha in comparison to the other alphas could be meaningful since, after all, this rater showed less internal consistency among his ratings.

To see which test rating given by the four raters agreed the least with the IELTS scores, intercorrelations of each test rating with the IELTS scores were computed as shown in Table 2. As the intercorrelations of the first rater demonstrate, Rater 1's CPE rating and Rater 4's TOEFL iBT rating show lower correlations with the IELTS ratings. Afterward, an alpha was computed for an aggregate of the ratings of all the raters including the IELTS scores.

Table 3 shows an alpha of around .86, which could be considered acceptable with regard to the small number of ratings.

To see which rating had a negative effect on the total alpha, item-total correlation for each test rating was computed. Item-total correlation showed the extent to which each test rating agrees with the total of the other test ratings including the IELTS scores. As it is shown in Table 4, CPE1 and iBT4 had the lowest correlations with the total ratings. This table also indicates that the removal of these scores would have increased the total alpha considerably.

These results, as expected, confirmed the results found in each rater's alpha and inter-test correlations computed in the previous section.

Factor analysis

This study was carried out having hypothesized that the construct of essay writing is similar across different standardized tests (i.e., IELTS, CAE, CPE, TOEFL iBT, and GRE), and a given essay is expected to be scored similarly by the rubrics and scales of these different exams. To see whether this was the case, the ratings of these exams were

Table 1 Reliability statistics

Raters	Cronbach's alpha based on standardized items	N of items
Rater 1	.67	5
Rater 2	.80	5
Rater 3	.80	5
Rater 4	.79	5

Table 2 Inter-item correlation matrix

		CAE	CPE	GRE	iBT
Rater 1	IELTS	.39	.23	.33	.46
Rater 2	IELTS	.48	.43	.47	.50
Rater 3	IELTS	.31	.38	.32	.48
Rater 4	IELTS	.53	.48	.50	.11

examined. The correlation analyses reported above showed that there is an acceptable agreement among all test ratings except two of them, CPE and TOEFL iBT. That is, rater 1 in CPE and Rater 4 in TOEFL iBT showed the least correlation among other test ratings (.15 and .13, respectively). To have a better picture of this issue, it was decided to run a PCA to examine the extent these exams tap the same underlying construct. Factor analysis provides some factor loadings for each test item (i.e., test rating); if two or more items load on the same factor, it will show that these items (i.e., test ratings) tap the same construct (i.e., essay writing construct).

Table 5 presents the results of Kaiser-Meyer-Olkin measure (KMO) and Bartlett's test of sphericity on the sampling adequacy for the analysis. The reported KMO is .83, which is larger than the acceptable value (KMO > .5) according to Field (2009). Bartlett's test of sphericity [χ^2 (136) = 1377.12, p < .001] was also found significant, indicating large enough correlations among the items for PCA; therefore, this sample could be considered adequate for running the PCA.

The next step was to investigate the number of factors required to be retained in the PCA. To do so, the scree plot was checked (Fig. 1). The first point that should be identified in the scree plot is the point of inflexion, that is, where the slopes of the line in the scree plot changes dramatically. Only those factors, which fall to the left of the point of inflexion, should be retained. Based on Fig. 1, it seems that the point of inflexion is on the fourth factor; therefore, four factors were retained.

According to Table 6, the first four retained factors explain around 60 percent of the whole variance, which is quite considerable.

Table 7 presents the four factor loadings after varimax rotation. Obviously, the different test ratings were loaded on 4 factors. In other words, those test ratings that clustered around the same factor seemed to be loading on the same underlying factor or latent variable.

Following the above analysis, it was decided to further examine the factor loadings as follows: It should be noted that the above factor structure was achieved by considering only those loadings above .4 as suggested by Stevens (2002), which explained around 16 percent of the variance in the variable. This value was strict, though, resulting in the emergence of limited factors. Therefore, employing Kaiser's criterion, a second factor analysis was run with a more lenient absolute value for each factor, which was .3 as suggested by Field (2009). By so doing, more factor loadings emerged and more information was achieved. The factor loadings above .3 are presented in Table 8, which almost revealed the same factor structure as found in the previous factor analysis with

Table 3 Reliability statistics

Cronbach's alpha based on standardized items	N of items
.86	17

Table 4 Item-total statistics

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
CAE1	72.38	111.04	.38	.67
CPE1	73.27	101.76	.15	.72
GRE1	73.06	114.00	.28	.68
iBT1	72.57	108.36	.47	.66
CAE2	72.49	110.94	.45	.67
CPE2	73.27	113.35	.44	.67
GRE2	72.86	111.91	.46	.67
iBT2	72.41	109.52	.46	.66
IELTS	71.18	109.45	.66	.66
CAE3	73.28	110.50	.36	.67
CPE3	73.59	112.63	.45	.67
GRE3	73.01	113.27	.41	.67
iBT3	73.32	107.66	.53	.66
CAE4	72.65	107.11	.57	.65
CPE4	73.11	108.90	.52	.66
GRE4	72.67	107.82	.57	.66
iBT4	72.40	80.45	.13	.80

absolute values greater than .4; however, one important finding was that the IELTS ratings this time showed loadings on all the factors on which other tests also loaded. It can be construed, therefore, that the other tests had significant potential to tap the same construct.

After estimating reliability using Cronbach's alpha and then by running a Confirmatory Factor Analysis, it was decided to omit Rater 1 due to his unfamiliarity with the exam and its corresponding rubrics reported by him in the questionnaire.

Table 9 and Fig. 2 (scree plot) demonstrate the factor structure after removing Rater 1. The scree plot shows that 4 factors should be retained in the analysis, and Table 9 indicates that the first four retained factors explain about 70 percent of the whole variance, which was quite satisfactory.

Finally, Table 10 shows that after removing Rater 1's data, all the ratings of Raters 3 and 4 have loaded on the same factors with the IELTS. Of course, like the previous factor analysis, the IELTS ratings again showed loadings on all the factors on which other tests loaded except iBT4. All in all, it could be concluded that the results from the factor analysis confirm the previous findings from alpha computations showing iBT4 ratings had the lowest correlations with the total ratings.

Table 5 KMO and Bartlett's test

Kaiser-Meyer-Olkin measure of sampling adequacy		**.83**
Bartlett's test of sphericity	Approx. chi-square	1377.12
	df	136
	Sig.	.001

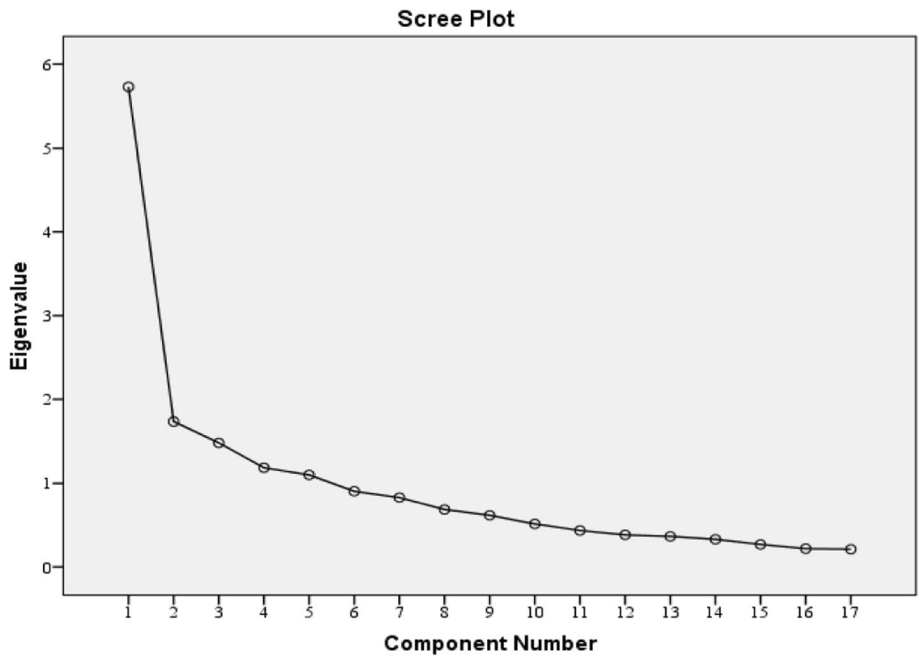

Fig. 1 Scree plot

Discussion and conclusions

The purpose of the present study is to examine the consistency of the rubrics endorsed for assessing the writing tasks by the internationally recognized tests of English language proficiency. Standard rubrics can be considered constructive tool helping raters to assess different types of essays (Busching, 1998). Using rubrics enhances the reliability of the assessment of essays provided that these rubrics are well described and that they tap the same construct (Jonsson & Svingby, 2007). The current study is an attempt to examine the reliability among different rubrics of essay writing with regard to their major components, namely, organization, coherence and cohesion, range of lexical and grammatical complexity used, and accuracy.

The results of this study show that all in all, there is a high correlation among raters (i.e., the IELTS examiner and the four other raters) and rating scores (i.e., the official IELTS scores and the other 16 test ratings received from the four raters). The intercorrelations among test ratings and the raters as well as the computation of inter-item correlations between each test rating and the IELTS scores revealed that CPE1 and iBT4 had the least agreement with the official IELTS ratings. Therefore, these low correlations were investigated in a follow-up study by giving the four raters a questionnaire including both open-ended and closed-ended questions. The raters' responses to the questionnaire denoted the extent to which they were familiar with each exam and their corresponding rubrics.

The responses of two of the raters, that is, Rater 1 in CPE and Rater 4 in TOEFL iBT, proved to be illuminating in explaining their performance. Rater 1 s' responses to the questionnaire showed that he had no teaching experience for CPE classes. However, his responses to other questions of the questionnaire indicated his familiarity with this exam and its writing essay scoring rubrics. The responses of Rater 4 revealed that she had no teaching experience for TOEFL iBT and no familiarity with the exam and its

Table 6 Total variance explained

Component	Initial Eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.72	33.69	33.69	5.72	33.69	33.69	2.80	16.47	16.47
2	1.73	10.22	43.91	1.73	10.22	43.91	2.76	16.24	32.72
3	1.48	8.72	52.64	1.48	8.72	52.64	2.64	15.55	48.27
4	1.18	6.98	59.62	1.18	6.98	59.62	1.93	11.35	59.62
5	1.10	6.47	66.10						
6	.90	5.32	71.43						
7	.83	4.88	76.31						
8	.68	4.03	80.35						
9	.61	3.62	83.97						
10	.51	3.02	86.99						
11	.43	2.55	89.55						
12	.38	2.25	91.81						
13	.36	2.14	93.95						
14	.33	1.94	95.89						
15	.26	1.57	97.47						
16	.21	1.28	98.75						
17	.21	1.24	100.00						

Extraction method: principal component analysis

Table 7 Rotated component matrix

	Component			
	1	2	3	4
CAE2	.83			
iBT2	.77			
CAE1	*.59*			
iBT1	*.51*			
IELTS	.51			
CAE3		.82		
iBT3		.77		
CPE3		.70		
GRE3		.66		
CAE4			.80	
GRE4			.78	
CPE4			.70	
iBT4			.51	
GRE1				*.68*
CPE1				*.64*
GRE2			.40	.64
CPE2				.48

Extraction method: principal component analysis
Rotation method: Varimax with Kaiser normalization

Table 8 Rotated component matrix[a]

	Component				
	1	2	3	4	5
CAE3	.82				
iBT3	.76				
CPE3	.75				
GRE3	.68				
CAE4		.80			
GRE4		.79			
CPE4		.74			
iBT4		.50			
CAE2			.82		
iBT2			.80		
IELTS			.40		
CAE1				.77	
iBT1				.73	
GRE1				.57	.53
GRE2					.69
CPE1					.66
CPE2			.42		.53

Extraction method: principal component analysis
Rotation method: Varimax with Kaiser normalization
[a]Rotation converged in 7 iterations, with 5 factors specified

Table 9 Total variance explained (Rater 1 removed)

Comp.	Initial Eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.05	38.89	38.89	5.05	38.89	38.89	2.70	20.83	20.83
2	1.71	13.16	52.06	1.71	13.16	52.06	2.65	20.39	41.23
3	1.31	10.13	62.19	1.31	10.13	62.19	2.56	19.70	60.93
4	.91	7.00	69.20	.91	7.00	69.20	1.07	8.27	69.20
5	.85	6.60	75.80						
6	.65	5.02	80.82						
7	.53	4.09	84.92						
8	.46	3.53	88.45						
9	.38	2.97	91.43						
10	.37	2.87	94.30						
11	.28	2.22	96.53						
12	.23	1.80	98.33						
13	.21	1.66	100						

Extraction method: principal component analysis

corresponding rating scales. The outcome from the interview with Rater 4 suggests that using well-trained raters leads to fewer problems in rating. What Rater 4 stated in her responses to the questionnaire and interview were in line with the findings of Sasaki and Hirose (1999), who concluded that familiarity with different tests and their relevant rubrics leads to better scoring. Additionally, the results of the present study are consistent with what Schoonen (2005), Attali and Burstein (2005), Wind et al. (2018), Deygers

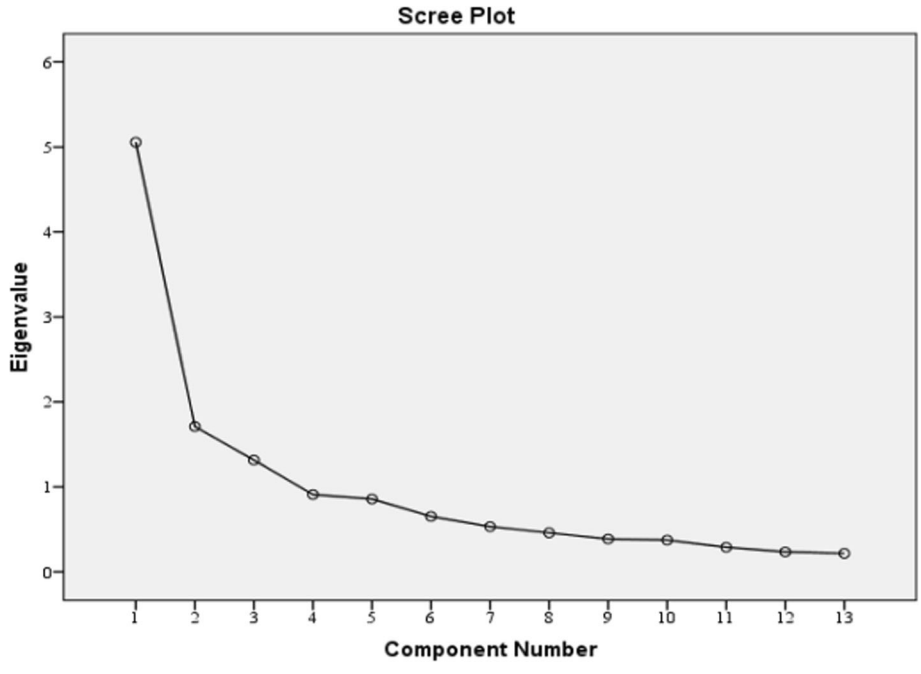

Fig. 2 Scree plot (Rater 1 removed)

Table 10 Rotated component matrix[a] (Rater 1 removed)

	Component			
	1	2	3	4
CAE3	.84			
iBT3	.79			
CPE3	.75			
GRE3	.68			
GRE2		.76		
CPE2		.74		
iBT2		.74		
CAE2		.63	−.37	−.37
IELTS	.34	.55	.44	
CAE4			.85	
CPE4			.83	
GRE4			.81	
iBT4				.79

Extraction method: principal component analysis
Rotation method: Varimax with Kaiser normalization
[a]Rotation converged in 6 iterations, with 4 factors specified

et al. (2018), Wesolowski et al. (2017), Trace et al. (2016), Fleckenstein et al. (2018), Rupp et al. (2019) found in their studies, that is employing rubrics enhances the reliability of writing assessment as well as among raters.

To this point, the obtained results from this study provide an affirmative answer to the first question of the study, indicating a very high agreement among test ratings and the raters. Also, in order to ensure that the construct of essay writing is similar across different standardized tests and identical essays are scored similarly by the internationally recognized rubrics of these different exams, inter-item correlation analysis was computed which indicated that CPE1 and iBT4 had the lowest correlations with the total ratings. This could be due to either the raters' inconsistencies or the hypothesis that essay writing is conceptualized differently based on the scoring rubrics of these exams. The follow-up survey also corroborated that the disagreement among Raters 1 and 4 and the other raters was due to either the rater's discrepancies or the way every writing task was hypothesized differently according to the rubrics of each exam. It can be supported by Weigle (2002) who concluded that raters should have a good grasp of scoring and its essential details. She also discussed that raters should have a sharp conceptualization of the construct of essay writing.

The results from the rotated component matrix revealed that all the ratings of Raters 3 and 4 loaded on the same factor, meaning that they tap the same construct. Examining the other factor loadings revealed that CAE1, iBT1, CAE2, and iBT2 also loaded on the same factor with the IELTS, suggesting that these rater's conceptualizations of the construct of essay writing in CAE and TOEFL iBT were more similar to that of the IELTS raters rather than those of CPE and GRE scorers. However, what remained questionable was why CPE1 and GRE1 did not load on the same factor as CAE1 and iBT1, and why CPE1 and GRE1 loaded on the same factor with CPE2 and GRE2. Additionally, why CPE1 also loaded with GRE2 and CPE2 on the same factor remained open to discussion.

What was found above was the results of the PCA considering those factor loadings above .4 based on Stevens (2002). As this value was strict, and the number of obtained factors was limited, it was decided to apply Kaiser's Criterion with a less rigorous eigenvalue of .3 based on Field' (2009) suggestion. The findings showed almost the same factor loadings as was found in the previous factor analysis. Again Raters 3 and 4 loaded on the same factor, but this time, the IELTS scores loaded on the same factor with CAE2 and iBT2. CAE1, GRE1, and iBT1 loaded on the same factor and what was still debatable was why CPE1 loaded with GRE 2 and CPE2.

Up to this mentioned point, all the results obtained from alpha computation and factor analysis indicated something different in Rater 1, based on which it was decided to omit Rater 1 from the PCA. It is interesting to note that after interviewing all the four raters and scrutinizing the questionnaire survey, it was found that Rater 1, in his responses to the questionnaire, had indicated that he had no teaching experience in teaching CPE classes, and yet he claimed that he was familiar with this exam and its related rating scales, contrary to other raters' responses to the questionnaire.

After omitting Rater 1 from the PCA, the findings showed that Rater 3's and Rater 4's test ratings loaded on the same factor, and this time the IELTS loaded on the factors that all the other tests had loaded except iBT4, meaning that Rater 4 had no agreement with the IELTS raters in rating the essay. What was found from the questionnaire survey of this rater indicated that Rater 4 had no teaching experience for this particular exam. She also had no familiarity with the exam and its corresponding rubrics. This rater also believed that scoring exams like TOEFL iBT and the exams developed by ETS were more difficult, and that they generally received lower scores in comparison to the Cambridge English Language Assessment exams. The results of what Rater 4 stated were not in line with the findings of Bachman (2000) who did a comparison study between TOEFL PBT and CPE essay task and concluded that CPE scoring is more difficult than scoring TOEFL PBT. Contrary to the findings of the present study, he also concluded that exams like CPE received lower scores.

The results from alpha computation and factor analysis showed the noticeable role of raters in assessing writing. The results from this study are in line with the findings of Lumley (2002) and Schoonen (2005) who argue that raters need to be considered one of the most remarkable concerns in the process of assessment. Shi (2001) argued in favor of the significant role of raters in assessing essays using their own criteria in addition to the standard and determined rating scales. Likewise, the outcome of factor analysis in this research study revealed that raters play a remarkable role in assessing essays by showing that all the items (i.e., test ratings) load on the same factor, especially when all the essay writings were rated by the same rater.

This study aimed to examine the consistency and reliability among different standard rubrics and rating scales used for assessing writing in the internationally recognized tests of English language proficiency. The results from alpha estimation provide evidence for a strong association among the raters and test ratings. Also, what has been found from the PCA indicate that these test ratings tap the same underlying construct. This study encourages employing practical rater trainer and rater training courses, providing them with the authentic opportunities to get familiar with different rubrics. This area requires more investigation on how raters themselves might affect the rating and how employing trained and certified raters can affect the process of rating. Test

administrators and developers are the other groups who benefit from the findings of this study, since, when argued that all the test ratings tap the same underlying construct and different essay writing rating scales can be predictors of each other, it would be practical for them to set standard essay writing rubrics which can be used for rating and assessing writing. Also, as the findings of the present study alluded, the developers of the writing rubrics for these tests may also take into stock the implication that there are critical constructs within writing that weigh more heavily when being assessed across standardized measures. Teachers and learners are other groups who benefit from the result of this research study. They might devote less time on describing all these rubrics with their descriptions stated in different words. Instead, they could spend more time on practicing writing and essay writing tasks.

The study tried to examine the reliability of analytic rubrics used in assessing the essay component of the following standardized examinations: IELTS, TOEFL iBT, CAE, CPE, and GRE. While the first four of the tests listed above are indeed English language proficiency examinations designed to assess language skills of English as a Second Language (ESL) learners, the last one (i.e. GRE) is intended for those seeking admission to graduate programs in the U.S., regardless of the first language background. GRE candidates are, at minimum, bachelor degree holders, most of whom are native speakers of English whose education was completed in the English language, while the minority are international applicants to U.S. universities' master's and Ph.D. programs from various language backgrounds. GRE writing task, in other words, is not intended for L2 English learners. Therefore, it seems that juxtaposing the GRE requirements for the writing task, which zero in on argumentation and critical thinking, with English language proficiency standards as measured by the other four tests can dilute the generalizability of the results particularly with reference to this particular exam, due to the divergent assessment purposes and intended candidate profiles for this test. Future researchers are encouraged to take heed of this limitation in the present study.

Acknowledgements
The authors would like to thank the reviewers for their fruitful comments. We would also like to thank the raters who kindly accepted to contribute to this study.

Authors' contributions
The authors made almost equal contributions to this manuscript, and both read and approved the final manuscript.

Authors' information
Enayat A. Shabani[1] (eshabani@tums.ac.ir) is a Ph.D. in TEFL and is currently the Chair of the Department of Foreign Languages at Tehran University of Medical Sciences (TUMS). His areas of research interest include language testing and assessment, and internationalization of higher education.
Jaleh Panahi[2] (Jaleh.panahi@gmail.com) holds an M.A. in TEFL. She has been teaching English for 12 years with the main focus of IELTS teaching and instruction. She is currently a part-time instructor at the Department of Foreign Languages, Tehran University of Medical Sciences. Her fields of research interest are language assessment, and language and cognition.

Availability of data and materials
The authors were provided with the data for research purposes. Sharing the data with a third party requires obtaining consent from the organization which provided the data. The materials are available in the article.

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

Received: 16 June 2020 Accepted: 3 September 2020
Published online: 26 September 2020

References
Aish, F., & Tomlinson, J. (2012). *Get ready for IELTS writing*. London: HarperCollins.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., & Roberts, T. (2011). Criteria for assessment: consensus statement and recommendations form the Ottawa 2010 conference. *Medical Teacher*, *33*(3), 206–214.
Anderson, C. (2005). *Assessing writers*. Portsmouth: Heinemann.
Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–18.
Archibald, A. (2001). Targeting L2 writing proficiencies: Instruction and areas of change in students' writing over time. *International Journal of English Studies*, *1*(2), 153–174.
Archibald, A. (2004). Writing in a second language. In *The higher education academy subject centre for languages, linguistics and area studies* Retrieved from http://www.llas.ac.uk/resources/gpg/2175.
Arter, J. A., Spandel, V., Culham, R., & Pollard, J. (1994). *The impact of training students to be self-assessors of writing*. New Orleans: Paper presented at the Annual Meeting of the American Educational Research Association.
Aryadoust, V., & Riazi, A. M. (2016). Role of assessment in second language writing research and pedagogy. *Educational Psychology*, *37*(1), 1–7.
Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater.V.2.0. (RR- 04-45)*. Princeton: ETS.
Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141.
Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell? *System*, *29*(3), 371–383.
Bachman, L., & Palmer, A. S. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
Bachman, L. F. (2000). Modern language testing at turn of the century: assuring that what we count counts. *Language Testing*, *17*(1), 1–42.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.
Bauer, B. A. (1981). *A study of the reliabilities and the cost-efficiencies of three methods of assessment for writing ability*. Champaign: University of Illinois.
Becker, A. (2011). Examining rubrics used to measure writing performance in U.S. intensive English programs. *The CATESOL Journal*, *22*(1), 113–117.
Bell, R. M., Comfort, K., Klein, S. P., McCarffey, D., Ormseth, T., Othman, A. R., & Stecher, B. M. (2009). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, *11*(2), 121–137.
Betsis, A., Haughton, L., & Mamas, L. (2012). *Succeed in the new Cambridge proficiency (CPE)- student's book with 8 practice tests*. Brighton: GlobalELT.
Biber, D., Byrd, M., Clark, V., Conrad, S. M., Cortes, E., Helt, V., & Urzua, A. (2004). Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus. In *ETS research report series (RM-04-3, TOEFL Report MS-25)*. Princeton: ETS.
Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university*. Maidenhead: McGraw Hill.
Birky, B. (2012). A good solution for assessment strategies. *A Journal for Physical and Sport Educators*, *25*(7), 19–21.
Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*(4), 410–421.
Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students. In *ETS Research Report Series (RR- 83-18, TOELF- RR-15)*. Princeton: ETS.
Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman, & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*, (pp. 112–140). Cambridge: Cambridge University Press.
Broad, B. (2003). *What we really value: beyond rubrics in teaching and assessing writing*. Logan: Utah State UP.
Broer, M., Lee, Y. W., Powers, D. E., & Rizavi, S. (2005). Ensuring the fairness of GRE writing prompts: Assessing differential difficulty. In *ETS research report series (GREB Report No. 02-07R, RR-05-11)*.
Brookhart, G., & Haines, S. (2009). *Complete CAE student's book with answers*. Cambridge: Cambridge University Press.
Brookhart, S. M. (1999). The art and science of classroom assessment: the missing part of pedagogy. *ASHE-ERIC Higher Education Report*, *27*(1), 1–128.
Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), *Writing assessment: issues and strategies*, (pp. 168–182). New York: Longman.
Brown, A., & Jaquith, P. (2007). *Online rater training: perceptions and performance*. Dubai: Paper presented at Current Trends in English Language Testing Conference (CTELT).
Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105–121.
Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practice*. Lewiston: Pearson Longman.
Brown, J. (2002). Training needs assessment: a must for developing an effective training program. *Sage Journal*, *31*(4), 569–578 https://doi.org/10.1177/009102600203100412.
Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing. Cambridge applied linguistics series*. Cambridge: Cambridge University Press.
Busching, B. (1998). Grading inquiry projects. *New Directions for Teaching and Learning*, (*74*), 89–96.
Canseco, G., & Byrd, P. (1989). Writing required in graduate courses in business administration. *TESOL Quarterly*, *23*(2), 305–316.
Capel, A., & Sharp, W. (2013). *Cambridge english objective proficiency*, (2nd ed.,). Cambridge: Cambridge University Press.
Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English*, *18*(1), 65–81.
Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, *18*(1), 80–98 https://doi.org/10.1177/0741088301018001004.
Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many facet models. *Journal of Applied Measurement*, *2*(4), 379–388.
Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.
Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100–108.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World English*, *14*(1), 99–115.
Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (2012). *The Cambridge guide to second language assessment*. New York: Cambridge University Press.
Corry, H. (1999). *Advanced writing with English in use: CAE*. Oxford: Oxford University Press.
Crusan, D. (2015). And then a miracle occurs: the use of computers to assess student writing. *International Journal of TESOL and Learning*, *4*(1), 20–33.
Cumming, A. (2001). Learning to write in a second language: two decades of research. *International Journal of English Studies*, *1*(2), 1–23.
Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8.
Cumming, A. H., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*, *ETS Research Report Series (RM-00-5; TOEFL-MS-18)*. Princeton: ETS.
Dass, B. (2014). *Adult & continuing professional education practices: CPE among professional providers*. Singapore: Partridge Singapore.
Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3–15 https://doi.org/10.1080/15434303.2016.1261350.
Diab, R., & Balaa, L. (2011). Developing detailed rubrics for assessing critique writing: impact on EFL university students' performance and attitudes. *TESOL Journal*, *2*(1), 52–72.
Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability (Research Bulletin No. RB-61-15)*. Princeton: Educational Testing Service https://doi.org/10.1002/j.2333-8504.1961.tb00286.x.
Dixon, N. (2015). *Band 9-IELTS writing task 2-real tests*. Oxford: Oxford University Press.
Duckworth, M., Gude, K., & Rogers, L. (2012). *Cambridge english: proficiency (CPE) masterclass: student's book*. Oxford: Oxford University Press.
Dunsmuir, S., & Clifford, V. (2003). Children's writing and the use of ICT. *Educational Psychology in Practice*, *19*(3), 171–187.
East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88–115.
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.
Eckes, T. (2012). Operational rater types in writing assessment: linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292.
Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64.
Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: does it work? *Language Assessment Quarterly*, *2*(3), 175–196.
Ene, E., & Kosobucki, V. (2016). Rubrics and corrective feedback in ESL writing: a longitudinal case study of an L2 writer. *Assessing Writing*, *30*, 3–20 https://doi.org/10.1016/j.asw.2016.06.003.
Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: a study of four experienced raters of ESL composition. In *ETS research report series (RR-03-17)*. Ontario: ETS.
Evans, V. (2005). *Entry tests CPE 2 for the revised Cambridge proficiency examination: Student's book*. New York City: Pearson Education.
Fahim, M., & Bijani, H. (2011). The effect of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1–16.
Faigley, L., Daly, J. A., & Witte, S. P. (1981). The role of writing apprehension in writing performance and competence. *Journal of Educational Research*, *75*(1), 16–21.
Field, A. P. (2009). *Discovering statistics using SPSS (and sex and drugs and rock' n' roll)*, (3rd ed.,). London: Sage Publication.
Fleckenstein, J., Keller, S., Kruger, M., Tannenbaum, R. J., & Köller, O. (2019). Linking TOEFL iBT writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43* https://doi.org/10.1016/j.asw.2019.100420.
Fleckenstein, J., Leucht, M., & Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly*, *15*(1), 90–101 https://doi.org/10.1080/15434303.2017.1421956.
Fleming, S., Golder, K., & Reeder, K. (2011). Determination of appropriate IELTS writing and speaking band scores for admission into two programs at a Canadian post-secondary polytechnic institution. *The Canadian Journal of Applied Linguistics*, *14*(1), *222–250*.
Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. New York: Routledge.
Gass, S., Myford, C., & Winke, P. (2011). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252.
Ghaffar, M. A., Khairallah, M., & Salloum, S. (2020). Co-constructed rubrics and assessment forlearning: The impact on middle school students' attitudes and writing skills. *Assessing Writing*, *45* https://doi.org/10.1016/j.asw.2020.100468.
Ghalib, T. K., & Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: a case study. *English Language Teaching*, *8*(7), 225–236.
Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: an applied linguistic perspective*. London: Longman.
Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: the effects of self-regulated strategy development. *Contemporary Educational Psychology*, *30*(2), 207–241 https://doi.org/10.1016/j.cedpsych.2004.08.001.
Gustilo, L., & Magno, C. (2015). Explaining L2 Writing performance through a chain of predictors: A SEM approach. *3 L: The Southeast Asian Journal of English Language Studies*, *21*(2), 115–130.
Hamp-Lyons, L. (1990). Second language writing assessment. In B. Kroll (Ed.), *Second language writing: research insights for the classroom*, (pp. 69–87). California: Cambridge University Press.
Hamp-Lyons, L. (1991). *Holistic writing assessment of LEP students*. Washington, DC: Paper presented at Symposium on limited English proficient student.

Hamp-Lyons, L. (2007). Editorial: worrying about rating. *Assessing Writing*, *12*, 1–9.
Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – writing: composition, community and assessment (toefl monograph series no. 5)*. Princeton: Educational Testing Service.
Harman, R. (2013). Literary intertextuality in genre-based pedagogies: building lexicon cohesion in fifth-grade L2 writing. *Journal of Second Language Writing*, *22*(2), 125–140.
Harrison, J. (2010). *Certificate of proficiency in English (CPE) test preparation course*. Oxford: Oxford University Press.
Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics*, *41*(4), 667–683.
Holmes, P. (2006). Problematizing intercultural communication competence in the pluricultural classroom: Chinese students in New Zealand University. *Journal of Language and Intercultural Communication*, *6*(1), 18–34.
Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.
Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–239.
Huot, B., Moore, C., & O'Neill, P. (2009). Creating a culture of assessment in writing programs and beyond. *College Composition and Communication*, ***61***(**1**), 107–132.
Hyland, K. (2004). *Disciplinary discourses: social interactions in academic writing*. Michigan: University of Michigan Press.
Inoue, A. (2004). Community-based assessment pedagogy. *Assessing Writing*, *9*(3), 208–238 https://doi.org/10.1016/j.asw.2004.12.001.
Izadpanah, M. A., Rakhshandehroo, F., & Mahmoudikia, M. (2014). On the consensus between holistic rating system and analytical rating system: a comparison between TOEFL iBT and Jacobs' et al. *composition. International Journal of Language Learning and Applied Linguistics World*, *6*(1), 170–187.
Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Rowley: Newbury House.
Jakeman, V. (2006). *Cambridge action plan for IELTS: academic module*. Cambridge: Cambridge University Press.
Jamieson, J., & Poonpon, K. (2013). Developing analytic rating guides for TOEFL iBT integrated speaking tasks. In *ETS research series (RR-13-13, TOEFLiBT-20)*. Princeton: ETS.
Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*, 121–138 https://doi.org/10.1207/S15324818AME1302_1.
Jones, C. (2001). The relationship between writing centers and improvement in writing ability: An assessment of the literature. *Journal of Education*, *122*(1), 3–20.
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.
Kane, T. S. (2000). *Oxford essential guide to writing*. New York: Berkey Publishing Group.
Kellogg, R. T., Turner, C. E., Whiteford, A. P., & Mertens, A. (2016). The role of working memory in planning and generating written sentences. *Journal of Writing Research*, *7*(3), 397–416.
Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparametrized unified model. *Language Testing*, *28*(4), 509–541.
Klein, P. D., & Boscolo, P. (2016). Trends in research on writing as a learning activity. *Journal of Writing Research*, *7*(3), 311–350 https://doi.org/10.17239/jowr-2016.07.3.01.
Knoch, U. (2009). The assessment of academic style in EAP writing: the case of the rating scale. *Melbourne Papers in Language Testing*, *13*(1), 35.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: what should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96.
Kondo-Brown, K. (2002). A facet analysis of rater bias in Japanese second language writing performance. *Language Testing*, *19*(1), 3–31.
Kong, N., Liu, O. L., Malloy, J., & Schedl, M. A. (2009). Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning. In *ETS research report series (RR-09-29, TOEFLiBT-09)*. Princeton: ETS.
Kroll, B. (1990). *Second language writing (Cambridge Applied Linguistics): research insights for the classroom*. Cambridge: Cambridge University Press.
Kroll, B., & Kruchten, P. (2003). *The rational unified process made essay: a practitioner's guide to the RUP*. Boston: Pearson Education.
Kuo, S. (2007). Which rubric is more suitable for NSS liberal studies? Analytic or holistic? *Educational Research Journal*, *22*(2), 179–199.
Lanteigne, B. (2017). Unscrambling jumbled sentences: an authentic task for English language assessment? *Studies in Second Language Learning and Teaching*, *7*(2), 251–273 https://doi.org/10.14746/ssllt.2017.7.2.5.
Leki, L., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. New York: Routledge.
Levin, P. (2009). *Write great essays*. London: McGraw-Hill Education.
Loughead, L. (2010). *IELTS practice exam: with audio CDs*. Hauppauge: Barron's Education Series.
Lu, J., & Zhang, Z. (2013). Assessing and supporting argumentation with online rubrics. *International Education Studies*, *6*(7), 66–77.
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: Lang.
MacDonald, S. (1994). *Professional academic writing in the humanities and social sciences*. Carbondale: Southern Illinois University Press.
Mackenzie, J. (2007). *Essay writing: teaching the basics from the group up*. Markham: Pembroke Publishers.
Malone, M. E., & Montee, M. (2014). *Stakeholders' beliefs about the TOEFL iBT test as a measure of academic language ability (TOEFL iBT Report No. 22, ETS Research Report No. RR-14-42)*. Princeton: Educational Testing Service https://doi.org/10.1002/ets2.12039.

Matsuda, P. K. (2002). Basic writing and second language writers: Toward an inclusive definition. *Journal of Basic Writing*, *22*(2), 67–89.
McLaren, S. (2006). *Essay writing made easy*. Sydney: Pascal Press.
McMillan, J. H. (2001). *Classroom assessment: principles and practice for effective instruction*, (2nd ed.,). Boston: Allyn & Bacon.
Melendy, G. A. (2008). Motivating writers: the power of choice. *Asian EFL Journal*, *20*(3), 187–198.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, *23*(2), 13–23.
Moore, J. (2009). *Common mistakes at proficiency and how to avoid them*. Cambridge: Cambridge University Press.
Moskal, B. M., & Leydens, J. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, *7*, 10.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12.
Muenz, T. A., Ouchi, B. Y., & Cole, J. C. (1999). Item analysis of written expression scoring systems from the PIAT-R and WIAT. *Psychology and Schools*, *36*(1), 31–40.
Muncie, J. (2002). Using written teacher feedback in EFL composition classes. *ELT Journal*, *54*(1), 47–53 https://doi.org/10.1093/elt/54.1.47.
Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.
Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, *14*(1), 3–26.
Nunn, R. C. (2000). Designing rating scales for small-group interaction. *ELT Journal*, *54*(2), 169–178.
Nunn, R. C., & Adamson, J. (2007). Toward the development of interactional criteria for journal paper evaluation. *Asian EFL Journal*, *9*(4), 205–228.
Nystrand, M., Greene, S., & Wiemelt, J. (1993). Where did composition studies come from? An intellectual history. *Written Communication*, *10*(3), 267–333.
O'Neil, T. R., & Lunz, M. E. (1996). *Examining the invariance of rater and project calibrations using a multi-facet rasch model*. New York: Paper presented at the Annual Meeting of the American Educational Research Associations.
Obee, B. (2005). *Practice tests for the revised CPE*. Berkshire: Express Publishing.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purpose revisited. *Educational Research Review*, *9*, 129–144.
Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: rasch partial credit analysis of performance in writing. *Language Testing*, *4*(1), 72–92.
Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly*, *25*(3), 407–430.
Reid, J. (1993). *Teaching ESL writing*. Englewood Cliffs: Regents Prentice Hall.
Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39.
Richards, J. C., & Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. New York: Pearson Education.
Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: recall, time and behavior observation training. *International Journal of Training and Development*, *7*(2), 93–107.
Rosenfeld, M., Courtney, R., & Fowles, M. (2004). *Identifying the writing tasks important for academic success at the undergraduate and graduate levels. Research report 42*. Princeton: Educational Testing Service.
Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *Identifying the reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels (TOEFL Monograph Series MS-21)*. Princeton: Educational Testing Service.
Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). *Automated essay scoring at scale: a case study in Switzerland and Germany (RR-86. ETS RR-19-12). ETS Research Report Series, 2019* https://doi.org/10.1002/ets2.12249.
Saeidi, M., Yousefi, M., & Baghayi, P. (2013). Rater bias in assessing Iranian EFL learners' writing performance. *Iranian Journal of Applied Linguistics*, *16*(1), 145–175.
Sasaki, M. (2000). Toward an empirical model of EFL writing processes: an explanatory study. *Journal of Second Language Writing*, *9*(3), 259–291.
Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, *16*(4), 457–478.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493.
Schirmer, B. R., & Bailey, J. (2000). Writing assessment rubric: an instructional approach for struggling writers. *Teaching Exceptional Children*, *33*(1), 52–58.
Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22*(1), 1–5.
Shaw, S. D., & Weir, C. J. (2007). *Examining writing: research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
Shermis, M. (2014). State-of-the-art automated essay scoring: competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76 https://doi.org/10.1016/j.asw.2013.04.001.
Shi, L. (2001). Native- and nonnative- speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303–325.
Spratt, M., & Taylor, L. B. (2000). *The Cambridge CAE course: self-study student's book*. Cambridge: Cambridge University Press.
Spurr, B. (2005). *Successful essay writing for senior high school*. NSW: New Frontier Publishing.
Staff, M. P. (2017). GRE guide to the use of scores. In *Graduate record examination*. Princeton: ETS.
Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*, (4th ed.,). Hillsdale: Erlbaum.
Stewart, A. (2009). *IELTS preparation & practice: reading and writing—academic module*. New York: Pearson Education.
Tardy, M. C., & Matsuda, P. K. (2009). The construction of author voice by editorial board members. *Written Communication*, *26*(1), 32–52.
Trace, J., Meier, V., & Janseen, G. (2016). "I can see that": developing shared rubric category interpretations through score negotiation. *Assessing Writing*, *30*, 32–43 https://doi.org/10.1016/j.asw.2016.08.001.

Ward, J. R., & McCotter, S. S. (2004). Reflection as a visible outcome for preservice teachers. *Teaching and Teacher Education*, *20*(3), 243–257.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*, 85–99.
Weir, C. J. (1990). *Communicative language testing*. New Jersey: Prentice Hall, Inc.
Weissberg, B. (2000). Developmental relationship in the acquisition of English syntax: Writing vs. speech. *Journal of Learning and Instruction*, *10*(1), 37–53 https://doi.org/10.1016/S0959-4752(99)00017-1.
Wesolowski, B. W., Wind, S. A., & Engelhard, G. (2017). Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, (*212*), 75–98 https://doi.org/10.5406/bulcouresmusedu.212.0075.
White, E. M. (1984). *Teaching and assessing writing*, (2nd ed.,). San Francisco: Jossey-Bass.
White, E. M. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass.
White, E. M. (1994). *Teaching and assessing writing*, (2nd ed.). San Francisco: Jossey-Bass.
Wiggan, G. (1994). The constant danger of sacrificing validity to reliability: making writing assessment serves writer. *Assessing Writing*, *1*, 129–139 https://doi.org/10.1016/1075-2935(94)90008-6.
Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Postmouth: Heinemann.
Wilson, M. (2017). *Reimaging writing assessment: from scales to stories*. Postmouth: Heinemann.
Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing*, *43* https://doi.org/10.1016/j.asw.2019.100416.
Wind, S. A., Tsai, C. L., Grajeda, S. B., & Bergin, C. (2018). Principals' use of rating scale categories in classroom observation for teacher evaluation. *School Effectiveness and School Improvement*, *29*(3), 485–510 https://doi.org/10.1080/09243453.2018.1470989.
Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, *2*(1), 59–61.
Wyldeck, K. (2008). *Everyday spelling and grammar*. Sydney: Pascal Press.
Zahler, K. A. (2011). *McGraw-Hill's conquering the NEW GRE verbal and writing*. New York: McGraw-Hill Education.
Zhang, B., Johnson, L., & Kilic, G. B. (2008). Assessing the reliability of self-and-peer rating in student group work. *Assessment & Evaluation in Higher Education*, *33*(3), 329–340 https://doi.org/10.1080/02602930701293181.

Publisher's Note