Assessing Writing 18 (2013) 111–131

Contents lists available at SciVerse ScienceDirect

Assessing Writing

Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays

Rajab Esfandiari [a,*], Carol M. Myford [b]

[a] Imam Khomeini International University, Department of English Language, Faculty of Humanities, Qazvin, Iran
[b] University of Illinois at Chicago, College of Education, Department of Educational Psychology, 1040 W. Harrison St. MC 147, Chicago, IL, 60607, USA

ARTICLE INFO

Article history:
Received 25 June 2012
Received in revised form 7 December 2012
Accepted 20 December 2012
Available online 7 February 2013

Keywords:
Writing assessment
Self-assessment
Peer-assessment
Rater effects
Rating scales

ABSTRACT

We compared three assessor types (self-assessors, peer-assessors, and teacher assessors) to determine whether they differed in the levels of severity they exercised when rating essays. We analyzed the ratings of 194 assessors who evaluated 188 essays that students enrolled in two state-run universities in Iran wrote. The assessors employed a 6-point analytic scale to provide ratings on 15 assessment criteria. The results of our analysis showed that of the three assessor types, teacher assessors were the most severe while self-assessors were the most lenient, although there was a great deal of variability in the levels of severity that assessors within each type exercised.

1. Introduction

The advent of alternative assessment in the early 1990s opened up new opportunities and horizons for language education, language classrooms, and language assessment. In the words of Hargreaves, Earl, and Schmidt (2002), these alternative assessment procedures

> motivate students to take more responsibility for their own learning, to make assessment an integral part of the learning experience, and to embed it in authentic activities that recognize and stimulate students' abilities to create and apply a wide range of knowledge, rather than simply engaging in acts of memorization and basic skill development. (p. 70)

* Corresponding author. Tel.: +98 914 1105945; fax: +98 281 8371602.
E-mail addresses: rbesfandiari@gmail.com (R. Esfandiari), cmyford@uic.edu (C.M. Myford).

1075-2935/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.asw.2012.12.002

Several innovative alternative assessment procedures, self-assessment and peer-assessment, are proving promising (Hamayan, 1995; Herman, Aschbacher, & Winters, 1992; Ross, 2005). In the next two sections, we describe these assessment practices, discuss their proposed benefits and limitations, and then summarize key findings from research on them.

1.1. Self-assessment

Researchers define self-assessment in a variety of different ways. For example, Andrade, Du, and Mycek (2010) defined self-assessment as "a process of formative assessment during which students reflect on the quality of their work, judge the degree to which it reflects explicitly stated goals or criteria, and revise accordingly" (p. 3). Viewed from this formative perspective, self-assessments are called "student evaluations." By contrast, viewed from a summative perspective, self-assessments are often referred to as "student grading" or "student marking," with teachers using the scores from such assessments when they assign course grades. More specifically, self-assessment refers to the "process in which students assess their own learning, particularly their achievements and learning outcomes" (Lindblom-Ylänne, Pihlajamaki, & Kotkas, 2006, p. 52).

Conceptually, self-assessment is supported by theories of cognition, constructivism and learner autonomy, especially those of Piaget and Vygotsky (Chen, 2008). According to these theorists, knowledge and meaning are constructed, with every individual having his/her own method of knowledge and meaning construction. Moreover, from these perspectives, "meaningful learning is reflective, constructive and self-regulated" (Maclellan, 2004, p. 311), and learners are required to think actively to construct mental models (Falchikov & Goldfinch, 2000). Self-assessment is also an essential aspect of autonomous and life-long learning (Jiliang & Kun, 2007). It derives its theoretical justification from principles of autonomy, intrinsic motivation, and cooperative learning (Brown, 2004). Through accurate appraisal of their own performance, students learn to rely on themselves, take responsibility for their own learning, and discern their own individual patterns of strengths and weaknesses, helping them to become more reflective, engaged learners. From a democratic perspective, assessment is no longer a unilateral teacher-centered activity (Shohamy, 2001). Rather, students work with the teacher to create criteria and standards, which in turn leads to greater student self-determination (Chen, 2008).

Self-assessment offers numerous possible advantages. In language teaching and learning, self-assessment can promote learning, raise learner awareness, aid in the establishment of learner goals, foster life-long learning, and facilitate democratic learning processes and needs analysis (Oscarson, 1989). Additionally, self-assessment can encourage greater effort, boost self-confidence, and facilitate awareness of the distinctions between competence and performance, as well as self-awareness of learners' strengths and weaknesses (Blue, 1994).

Despite the advantages, self-assessment also has certain limitations. One concern is some students' tendency to overestimate their capabilities and performance, hence inflating their scores and ratings (Lindblom-Ylänne et al., 2006). According to Evans, McKenna, and Oliver (2005), there are three reasons that may explain why self-assessors tend to rate more leniently than other assessor groups: (a) they may misunderstand what is expected if they have not been actively involved in the creation of the assessment tool, (b) they may deceive themselves about their abilities, or (c) they may score based on their views of their potential rather than on their actual abilities. Additionally, self-assessors may not have clear knowledge and understanding of the assessment criteria if they are not well trained or have not received explicit instruction regarding how to use these criteria (Leach, 2012). However, Ross and Starling (2008) have shown that if students know that their instructors will be reviewing their ratings, then students are more likely to try to be as accurate as possible. Kirby and Downs (2008) attributed student self-assessors' overrating to their belief that they should receive higher ratings not only because they have worked hard and exerted much effort, but also because they believe they have demonstrated that they have met the assessment criteria.

Closely related to the problem of overestimation is that of the accuracy of self-assessments, which can influence the reliability and validity of self-ratings. For example, Topping (2003) noted that grades that students assigned to themselves tended to be higher than grades that teachers assigned to them, and that self-assessments that were based on the students' perceptions of their levels of effort rather than their levels of achievement were particularly unreliable. Furthermore, self-assessments tend to

be more unreliable when students assess their own performances than when they assess their own learning products (Segers & Dochy, 2001).

Students' feelings and attitudes toward self-assessment may also prove problematic, exerting an influence on the ratings they assign. In some instances, students fear being wrong in their self-assessments and thus are reluctant to want to evaluate their own work (Leach, 2012). Additionally, some students harbor negative attitudes toward self-assessment. They dislike assessing themselves or their peers and would much prefer to have teachers assess them (Evans et al., 2005).

Researchers have conducted a number of studies of self-assessment in various content areas, language learning, and L2 writing. In a meta-analysis of 48 studies carried out in higher education settings, Falchikov and Boud (1989) identified three factors in studies that showed a close correspondence between student self-assessment ratings and teachers' ratings of those students: the quality of the study design, the level of the course, and the content area. They reported that (a) students and teachers rated more accurately in the better designed studies, (b) students in advanced courses rated more accurately than students in introductory courses, and (c) students in science courses rated more accurately than students in other courses. In a meta-analytic study of self-assessment in second/foreign language learning settings, Blanche and Merino (1989) summarized the L2 literature on self-assessment of foreign language skills. The researchers drew the following conclusions: (a) students' self-assessment ratings showed varying degrees of correspondence with measures of different external criteria; (b) the accuracy of most students' self-estimates was somewhat variable, depending upon their linguistic skills and what it was that students were rating, with estimates being generally accurate or very accurate; and (c) more proficient students tended to underrate themselves, while less proficient students tended to overrate themselves. In another meta-analysis of 11 correlational studies carried out in L2 language testing settings, Ross (1998) found robust concurrent validity evidence for students' self-assessment ratings: there were positive correlations between those ratings and measures of other language skills-related criterion variables such as writing, reading, speaking and listening. He concluded that the amount of prior experience that learners had in self-assessment contexts could affect the accuracy of their ratings.

Ross (2006) reviewed studies investigating the reliability, validity, and utility of self-assessment across diverse L1 and L2 settings. The studies included various types of items, tasks and learners. With regard to the reliability of self-assessment ratings, he concluded that (a) scores on self-assessments typically showed high internal consistency across items and tasks, particularly when they were administered over short time periods; (b) scores on self-assessments were less reliable when young learners were involved; and (c) when learners were trained to self-assess, adequate consistency was achieved. As for the validity of self-assessments, among the major findings were that (a) students' assessments of their own performances tended to be higher than their teachers' assessments of them; (b) students tended to overestimate their abilities when the scores were supposed to contribute to their grades; (c) there was higher student–teacher agreement when the students were taught how to assess their work, when they had sufficient knowledge of the content of the domain in which the task was embedded, when they knew their self-assessments were to be compared with their peers' or teachers' ratings, and/or when they applied assessment criteria that involved making low-level inferences; and (d) there was generally higher self-peer agreement than self-teacher agreement.

1.2. Peer-assessment

In broad terms, "peer-assessment involves collaboration in the appraisal of learning outcomes by those involved in the learning process, i.e., students" (van Gennip, Segers, & Tillema, 2009, p. 41). More clearly and unambiguously, "peer-assessment is the process through which groups or individuals rate their peers" (Falchikov, 1995, p. 175). Even more explicitly, peer-assessment is "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (Topping, 2010, p. 62).

Although peer-assessment and peer feedback are sometimes used interchangeably in the literature (see Gielen, Peeters, Dochy, Onghena, & Struyven, 2010), peer-assessment differs markedly from peer feedback, which is usually "an activity frequently used in second/subsequent (L2) writing classrooms to elicit feedback from a sympathetic reader, another student writer, on a draft version of a

text" (McGarrell, 2010, p. 72), "a communication process through which learners enter into diologue related to performance and standards" (Liu & Carless, 2006, p. 280),"often associated with process approaches to writing instruction" (Jacobs, Curtis, Braine, & Huang, 1998, p. 308). Accordingly, "peer feedback is primarily about rich detailed comments but without formal grades, whilst peer-assessment denotes grading (irrespective of whether comments are also included)" (Liu & Carless, 2006, p. 280). By implication, peer-assessment is quantitative, and peer feedback is qualitative.

According to Pope (2001) and Weaver and Esposto (2012), peer-assessment can take three different forms:

- Peer nomination, the act of identifying the best and the worst performers in the group,
- Peer rating, assessing each individual performer based on a set of performance or assessment criteria, or
- Peer rankings, ranking individual performers from the best to the worst against a set of criteria.

As Pope (2001) claimed, "these forms, as mechanical devices, sit well with the practices [of] ... group assessment; peer feedback and self-assessment; assessment of process and, negotiated assessment" (pp. 236–237). Allowing students to grade each other offers four potential benefits over teacher grading: logistical (i.e., teachers' time would be saved, and peers could provide quicker feedback than teachers); pedagogical (i.e., students could deepen their understanding by reading their peers' comments); metacognitive (i.e., students could be made aware of their strengths and weaknesses and could take responsibility for their own learning); and affective (i.e., classes could become more productive, friendlier, and more cooperative) (Sadler & Good, 2006).

From a pedagogical point of view, peer-assessment can promote student learning (Falchikov & Goldfinch, 2000) through "a sense of ownership and responsibility, motivation, and reflection of the student's own learning" (Saito & Fujita, 2009, p. 151). From an educational perspective, peer-assessment can provide other students with detailed peer feedback (Topping, 2009). Peer-assessment can engage students in the joint construction of knowledge through collaboration and mutual discourse and can develop students' higher order thinking and reasoning processes (Cheng & Warren, 2005).

Like self-assessment, peer-assessment also has certain limitations. Some researchers have reported that the ratings of the peer-assessors in their studies were not as reliable, accurate, and/or precise as the ratings of teachers, rendering peer ratings to be of limited value for formal summative assessment purposes, in their judgment (Freeman, 1995; Goldfinch & Raeside, 1990; Kwan & Leung, 1996; Orsmond, Merry, & Reiling, 1996).

Ethical challenges may arise. For example, if the racial/ethnic backgrounds of a peer-assessor and his/her classmate differ, the classmate might claim that the peer-assessor's ratings are biased, possibly creating significant tensions between those two students. In some situations, a peer-assessor may feel very uncomfortable assigning low ratings, even if the peer-assessor believes that such ratings are warranted, given the quality of his/her classmate's work. The peer-assessor might worry about causing friction and hurt feelings, or, if the classmate is a friend, believing that he/she is betraying the friend's trust (Vu & Dall'Alba, 2007). Additionally, there may be instances in which some students may not take seriously their responsibility for rating their classmates' performances or products, which might lead to skepticism regarding the meaningfulness of peer-assessments (Freeman, 1995).

In some assessment settings, students may have difficulty understanding the criteria that they are to use to rate, especially if they do not receive training in how to apply the criteria (Orsmond et al., 1996), along with guided practice in scoring carefully chosen examples. This requires that teachers build into their schedules sufficient time to prepare, train, and monitor students so that students can carry out peer-assessment in a credible manner, which can appreciably increase teachers' (and students') workloads (Vu & Dall'Alba, 2007).

Some students may feel unqualified to rate one another's work and thus may be reluctant to do so (Orsmond et al., 1996), perhaps feeling psychologically unprepared (Mok, 2011). Students may be used to teachers taking responsibility for assessment and may be reluctant and uncertain about shouldering that responsibility themselves (Vu & Dall'Alba, 2007). Finally, students may be willing to

have certain classmates rate their work, but not other classmates, particularly if they do not respect those classmates (Evans et al., 2005).

Studies of peer-assessment have focused on different aspects of its utility. Some studies have examined the reliability and validity of peer-assessment. For example, Falchikov and Goldfinch (2000) conducted a meta-analysis of 48 peer-assessment studies carried out in higher education settings. In each of these studies, researchers compared peer and teacher marks. The results of their analysis showed that the correlations between the two sets of marks were higher in studies in which peer assessments involved making overall global judgments using well-understood criteria (as compared to studies in which teachers and students marked analytically, making judgments about several individual dimensions). The level of agreement between peer and teacher marks was not appreciably higher in studies that were carried out in advanced courses (as compared to those carried out in introductory courses). Additionally, there were no clear differences in levels of agreement attained in studies conducted in different content areas. Earlier, Topping (1998) concluded that the results from the majority of studies he reviewed that were carried out in higher education settings suggested that peer-assessment demonstrated adequate reliability and validity in a wide variety of applications, reinforcing the previous findings of Percival and Ellington (1984) who, in their review of research conducted in similar settings, concluded that peer-assessment was valid, reliable, practicable, fair, and useful to students.

Saito and Fujita (2004) succinctly summarized the possible values and benefits of self-assessment and peer-assessment in L2 settings as "(1) feedback from multiple perspectives, (2) increasing responsibility for managing the assessment process, (3) sensitiz[ing] students to the evaluation criteria, and (4) sharing about individual strengths and weaknesses" (pp. 34–35). As Saito (2008) noted, both self-assessment and peer-assessment are supported by "evidence from second language (L2) acquisition research, mainstream education research, and both L2 writing and L1writing research" (p. 553).

1.3. The possible role of culture in carrying out self-assessment and peer-assessment practices

While self-assessment and peer-assessment practices have the potential to motivate students and stimulate learning, some researchers contend that students' cultural backgrounds may influence how they carry out these practices. That is, cultural mores and expectations may play a role in determining how students perform these rating tasks. In this section, we present key findings from those studies.

In a study involving university students from such diverse countries as Bangladesh, Germany, Italy, Lebanon, Saudi Arabia, Syria and Tanzania who were enrolled in an ESL class (i.e., an English for Academic Purposes course), Blue (1994) compared students' self-assessments of ten language skill areas to their teacher's assessments of their language skills. The students rated their skill levels at the beginning of the course and then again six months later at the end of the course. The results showed that students had great difficulty assessing themselves, even with teacher feedback. At the beginning of the course, the correlations between the students' ratings of their language skills and their TOEFL and IELTS scores were low. Additionally, at the end of the course, the correlations between the students' ratings of their language skills and the teacher's ratings of their skills were low. In a table that identified each student by his/her country of origin, Blue compared the teachers' rank orderings of students by their speaking ability to the students' self-assessments. In some cases, he noted stark differences between the teachers' and the students' assessments. One possible explanation for these results, Blue argued, was that the students' nationalities and their cultural values might account for those differences, "with some nationalities having a tendency to overestimate their [language skill] level and others tending towards underestimation" (p. 15).

Chen (2008) investigated university students' oral performance in an EFL class in China. Although the results showed that students made significant progress in learning to assess their own oral performance, they were likely to "under-mark" themselves. Chen attributed this under-marking to Chinese culture, which emphasizes strict discipline, generosity to others, and modesty. These factors could have influenced students' self-assessments, causing them to assign lower scores than they deserved, he posited.

Matsuno (2009) studied Japanese students enrolled in university EFL writing classes, and his findings mirror those of Chen. The students tended to assign their own essays lower ratings than warranted,

while they tended to assign higher ratings to their peers' essays. Matsuno explained that in Japan, self-assessors tend to be critical of their writing abilities because it reflects the mores of their culture (i.e., students show their modesty by not exaggerating their abilities). Therefore, students will tend to assign ratings to their own essays that are lower than they actually deserve. Similarly, Brown (2005) found that the Japanese self-assessor in her small-scale study tended to underestimate her own writing ability while overestimating the writing abilities of her peers (i.e., third and fourth year undergraduate learners of Chinese, five who were of Anglo background, one who was Australian-born but of Chinese background, and one who was Chinese Mainland-born). Brown hypothesized that the Japanese self-assessor's tendency to be overly critical of her abilities may reflect "social and/or cultural factors coming into play." The student stated that she "struggled with modesty and ego when self-assessing" (p. 183). Reflecting on her findings, Brown noted that "self- and peer-assessment in contexts where those assessments will be shared with other people . . . is a social act, and therefore responsive to social constraints such as 'modesty and ego'" (p. 186).

Employing a qualitative methodology in an L2 setting, Leach (2000) interviewed 25 international students to learn about their experiences with self-directed learning in both formal and informal settings. Her sample consisted of three students who were Pacific Islanders, two who were Asian, and twenty who were Pākehā (i.e., New Zealanders of European descent). In her interviews, she included questions about the students' experiences with, and reactions to, self-assessment. One of the themes that emerged from her analysis of the transcripts of the interviews was that cultural values could impact students' abilities to accurately self-assess. For example, she noted that one student she interviewed stated that if he assigned himself high scores, other people within his culture might think that he was boasting.

1.4. Studies of differences in the severity of self-assessors, peer-assessors, and teacher assessors

While many researchers have examined the degree of correspondence between the ratings of self-assessors and teacher assessors (or between the ratings of peer-assessors and teacher assessors), only a few researchers have directly compared these assessor types to determine whether some tend to rate more severely or leniently than other assessor types. In this next section, we examine the few studies that have compared assessor types in terms of their levels of severity.

Nakamura (2002) studied peer-assessment and teacher assessment in an EFL setting, Japan. In this small-scale study, Nakamura analyzed the ratings that five student assessors and one teacher assessor assigned to the presentations of 12 students in an oral presentation course. The raters used rating scales to evaluate the students' presentations. He reported that (a) the teacher assessor was more lenient than the peer-assessors were, but the peer-assessors varied in their levels of severity and leniency; (b) peer-assessment was successful in motivating students to improve their oral skills; and (c) students as peer-assessors could be reasonably reliable raters of their peers.

Investigating characteristics of peer rating in Japan, Saito and Fujita (2004) studied self, peer, and teacher ratings of business management students' writing samples. The researchers asked 92 students (47 peer-raters and 45 self-raters) enrolled in a basic EFL writing course and two experienced teachers to rate the samples. They used a translated version of the ESL Composition Profile that Jacobs and his colleagues (1981) created to evaluate the students' writings. The results from their analysis showed that there was much more variability among the self-raters in the levels of severity they exercised than among the peer-raters or the teacher raters (i.e., the self-raters included not only the most lenient rater in the entire sample but also the most severe rater). The peer-raters were relatively lenient on average, compared to the other two groups. Additionally, the teacher raters were on average more severe than the peer-raters, but less severe than the self-raters. The researchers also found that there was a strong positive correlation between the peer ratings and the teacher ratings.

When Saito (2008) compared the ratings of peer-assessors and teacher assessors who evaluated the oral presentations of Japanese students in an EFL class, his results were similar to those of Saito and Fujita (2004). Saito analyzed the ratings that 74 first-year university students who were majoring in economics assigned, as well as the ratings that three teacher assessors assigned. He found that the teacher assessors were more severe than all groups of peer-assessors when assessing all aspects of the oral presentations except the visual aspect. Later, Saito and Fujita (2009) examined the similarities and

differences between instructors' and peer-assessors' ratings of EFL group presentations in a yearlong English course and found that there was a moderate-to-high positive correlation between their ratings (i.e., .74, $p < .01$).

Matsuno (2009) analyzed the ratings of 91 students and four teacher raters. He compared the ratings of self-, peer-, and teacher raters in university writing classes in Japan. He conducted a rater–writer bias interaction analysis and found that "self-raters tended to assess their own writing more strictly than expected" (p. 91). Moreover, in this study "high-achieving writers did not often rate their peers severely and low-achieving writers did not often rate their peers leniently" (p. 92), but peer-assessors showed "reasoned assessments independent of their own performances" (p. 92).

1.5. Studies of the effectiveness of rater training procedures

When raters are called upon to evaluate essays in high-stakes settings, they often participate in a rater training program. Weigle (2002) described rater training processes for large-scale writing assessment and noted that those processes could be adapted for other assessment settings, although she cautioned that such processes are context-dependent, varying depending on circumstances.

The first step in training raters frequently involves selecting a series of anchor/benchmark essays that illustrate the scoring characteristics at each score point on the rating scale. Each anchor/benchmark essay will typically be annotated, describing how the qualities of writing in the essay correspond to one of the score points on the scale. Next, the rater trainees will examine and discuss these essays in order to become familiar with the rating criteria, the scale, and the assessment procedure. The third step involves having the trainees practice rating some sample essays. Typically they will then discuss the ratings they assigned to clear up any misconceptions or misunderstandings they may have about how they are to use the various points on the scale. The goal of this task is for the trainees to develop a common understanding of the rating criteria and of the meaning of each point on the rating scale. The trainees will then rate another set of sample essays that experienced raters have previously rated. The trainers will compare the trainees' ratings to those that the experienced raters assigned to determine each trainee's level of accuracy. The purpose of this step is to certify trainees who have demonstrated that they can use the scale in a reliable, accurate manner and to identify any trainees who are in need of additional training. Finally, qualified trainees will be invited to participate in the operational scoring, in which they will independently evaluate each essay to assign their ratings (McNamara, 1996).

Researchers have examined the effects of rater training on subsequent rater performance and have reported some interesting results. Weigle (1998) employed sixteen raters (divided into two groups of eight raters whom she labeled New and Old raters) to rate compositions that college students wrote. The raters attended a 90-minute rater training session. She compared the raters' ratings before and after training to determine whether there were any differences in their levels of severity or consistency. Weigle reported that before training, the New raters were not only more severe, but also less consistent, than the Old raters. After training, the raters still showed significant differences in the levels of severity they exercised, but they had become more consistent. Weigle concluded that "rater training cannot make raters into duplicates of each other, but it can make raters more self-consistent" (p. 281).

Other researchers have reported similar results. For example, Knoch, Read, and von Randow (2007) compared the effectiveness of online and face-to-face rater training in the context of a large-scale academic writing assessment for students in an L2 setting in New Zealand. They reported that both training modes were effective, although online training was slightly more effective than face-to-face training in reducing severity/leniency differences between raters. Both rater groups rated in a consistent manner after training, although raters who participated in the online training tended to rate more consistently than raters who participated in the face-to-face training. Additionally, other researchers (Elder, Knoch, Barkhuizen, & von Randow, 2005; Lumley & McNamara, 1995; McIntyre, 1993; McNamara, 1996; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994) have reported results that suggest that rater training can reduce but not eliminate differences in rater severity, help raters to become more self-consistent, and reduce some biases that individual raters may display.

2. Rationale for the study

The results from prior research on self-assessment and peer-assessment that we have reported indicate that in some settings students are able to evaluate themselves and their peers accurately alongside their teachers. However, few researchers have investigated self-assessment and peer-assessment of L2 writing in EFL settings. Most often researchers have conducted their studies in L1 settings for formative assessment purposes. Additionally, most of these studies were correlational, examining the overall degree of correspondence between the ratings of self-assessors and teacher assessors (or between the ratings of peer-assessors and teacher assessors). Comparatively few researchers have investigated severity differences among self-assessors, peer-assessors, and teacher assessors. Methodologically, this is problematic since it is possible for two sets of ratings to show a high degree of correspondence, but for assessors in those two groups to differ markedly in the levels of severity they exercise. High inter-rater reliability does not necessarily signal assessor interchangeability. There is a need then for research that directly compares the severity of these assessor types as they assign ratings in EFL settings. It may be that with adequate training and supervision, students can learn to rate accurately their own writing products and/or those of their peers, exercising levels of severity similar to those of their teachers. If that were the case, then it would be feasible to include their ratings, alongside those of their teachers, when calculating students' final grades in courses (i.e., formally introducing their assessments into higher education for summative purposes).

The purpose of our study was to determine to what extent the ratings that self-assessors and peer-assessors assigned were similar to (or different from) the ratings that teachers assigned. In this study, we compared the ratings of peer-assessors, self-assessors, and teachers assessors to determine (a) whether some assessor types tended on average to rate more severely than other assessor types, and (b) whether some assessor types showed more variability in the levels of severity they exercised than other assessor types.

3. Research questions

We were interested in whether three assessor types, namely, self-assessors, peer-assessors, and teacher assessors, would differ in the average levels of severity they exercised when rating college students' L2 essays, and whether some assessor types would show more variability among their measures of severity than other assessor types. Thus, we posed the following two research questions to focus our study:

1. To what extent do three assessor types (teacher assessors, peer-assessors and self-assessors) differ in the average levels of severity they exercise when rating college students' L2 essays?
2. How variable are self-assessors, peer-assessors and teacher assessors in the levels of severity they exercise?

4. Method

4.1. Participants

The participants consisted of 194 students and teachers, who functioned as self-assessors, peer-assessors, and teacher assessors. The student assessors were 188 undergraduate Iranian English majors enrolled in Advanced Writing classes in two state-run universities in Iran, in three fields of study: English Literature, Translation Studies, and English Language Teaching. The student assessors functioned both as self-assessors and as peer-assessors. The teacher assessors were six Iranian teachers of English.

The student assessors ranged in age from 18 to 29, with one over 30, and another who did not indicate his/her age. One hundred and thirty-one student assessors (69.7%) were female and 57 (30.3%) were male. One hundred and five (55.85%) were native Farsi-speakers, 68 (36.17%) were native-Turkish speakers, 11 (5.85%) were native-Kurdish speakers and another four (2.13%) were grouped as "Other." Ninety-five (50.5%) were second-year students, 29 (15.5%) were third-year students, and 64 (34.0%)

were fourth-year students. Only three of them (1.6%) had experience living in an English-speaking country. The number of years they had studied English ranged from 1 to 24 years, and most of them (61.7%) had studied the English language in language institutes before entering the university.

The teacher assessors were all male. They came from two language backgrounds: four teacher assessors were native-Farsi speakers, and the other two were native-Turkish speakers. They ranged in age from 23 to 36. None of them had experience living in an English-speaking country. They had taught writing courses from one to seven years. Three of them were affiliated with a national university, one of them with a private university, and two of them were classified as "Other." All of them had a degree in English: three of them were PhD students in ELT, two had MAs in ELT, and one had a BA in English literature.

4.2. The assessment instrument

We chose to create an analytic assessment instrument (rather than a holistic assessment instrument) to evaluate students' essays for several reasons: (a) analytic rating scales can provide diagnostically useful information regarding the strengths and weaknesses of students' writing; and (b) once raters have been properly trained, raters can more easily understand and reliably apply analytic writing scales (Weigle, 2002).

We drew upon the components of writing identified in the ESL Composition Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981) to help us create our instrument, but our instrument differed from the ESL Composition Profile in several important respects (see Appendix A). Our instrument contained fifteen criteria that represent the elemental features of a high quality five-paragraph essay: substance, thesis development, topic relevance, introduction, coherent support, conclusion, logical sequencing, range, word choice, word form, sentence variety, overall grammar, spelling, essay format, and punctuation, capitalization, and handwriting. By contrast, the ESL Composition Profile identifies five components of writing that raters are to score—content, organization, vocabulary, language use, and mechanics—with several subcomponents embedded within each of those components, and the components are differentially weighted to arrive at a composite score.

The ESL Composition Profile uses 4-category scales for rating each of the five dimensions (i.e., very poor, fair to poor, good to average, excellent to very good), along with dimension-specific descriptors for each of the rating scale categories. By contrast, we employed a 6-category scale for rating each of our criteria because, as Schaefer (2008) has pointed out:

> These are the most common number of scale steps in college writing tests, and a larger number of steps may provide a degree of step separation difficult to achieve as well as placing too great a cognitive burden on raters, while a lower number may not allow for enough variation among the multifaceted elements of writing skills. (p. 473)

Our scale categories were very poor (1), poor (2), fair (3), good (4), very good (5), and excellent (6).

4.3. Data collection procedures

We collected 188 five-paragraph essays over a span of a year and a half from 188 students enrolled in Advanced Writing courses in two state-run prestigious universities in two different cities in Iran. The students came from six classes taught by four instructors. The students in Advanced Writing classes are taught punctuation, expression, features of a well-written paragraph, and principles for writing one-paragraph and five-paragraph essays. This is the overall format for an advanced writing course as set by the Ministry of Sciences, Research, and Technology in Iran, and all instructors must follow this syllabus. The students were taught these principles of writing in eight weekly meetings. Immediately after the eight meetings, the teachers told their students that they would have to take the midterm exam the following week.

During the exam, students had 90 minutes to write a five-paragraph essay ranging in length from 500 to 700 words on the following topic: In your opinion, what is the best way to choose a marriage partner? Use specific reasons and examples why you think this approach is the best. The lead researcher

chose this topic from a list of TOEFL TWE (Test of Written English) topics. All the students wrote to the same topic in order to control for a topic effect.

4.4. Assessor training

All the student assessors and teacher assessors participated in a one-hour training session, in which they were instructed regarding how to rate the essays. In both universities, the student assessors were in intact classes. First, the lead researcher talked to the teachers who were teaching the classes. When the lead researcher had the teachers' permission, he explained the purpose of the research to the students in each class. They were each given a consent form to sign. The lead researcher conducted two separate training sessions, one for students enrolled in each of the two universities. The training procedures included instructions regarding how to function as both self-assessors and peer-assessors.

The lead researcher gave students an essay rating sheet, an essay that a teacher had previously rated (with corrections included), and a set of written guidelines in Farsi explaining how to use the rating scale (see Appendix B). The lead researcher went over the guidelines orally with the students, discussing each of the criteria that students were to consider when assessing the essay. The lead researcher told them to read the rated essay first without paying attention to the corrections made on the essay. After they had finished reading the essay, the lead researcher directed their attention to the corrections that the teacher had made on the essay and to the essay rating sheet, noting particularly the ratings that the teacher had assigned the essay. The lead researcher then explained how to use the essay rating sheet and discussed the reasoning the teacher had employed when assigning ratings to that essay.

For practice, the lead researcher gave them a new essay that one of the students had written and told them to read the essay and assess it according to the guidelines. The lead researcher instructed them to follow closely the written guidelines. During this phase of the training, the lead researcher monitored their ratings and explained any unclear points.

Following the training session, each self-assessor rated his/her own essay. The lead researcher gave the students a new essay rating sheet, the set of written guidelines for conducting a self-assessment, and their own essays. They knew that they were to self-assess their own essays, and the lead researcher advised them to assess as accurately as possible. When they had completed this task, the researcher gave each student one of his/her classmates' essays, with names removed, to evaluate as a peer-assessor. The students used the same rating procedure when they were assessing their peers' essays. After they finished rating the essays, the lead researcher collected the rating sheets. The entire training and rating session took about two hours.

The lead researcher used the same training procedure when working with the teacher assessors. Since it was not possible to arrange for a group meeting, the researcher met the teachers individually, provided instruction in how to use the rating criteria, gave them all 188 essays to assess, and asked them to complete and submit their rating sheets in one month.

The teacher assessors involved in this study were not the instructors who were teaching the Advanced Writing classes. The student assessors and the teacher assessors rated the essays independently of one another. The lead researcher was present when students were rating their essays to monitor their rating, but it was not possible to monitor the rating of the teacher assessors, given their varied work schedules. Consequently, they were given a month to complete the rating of the essays.

5. Results

5.1. Statistical analysis procedures

We used a many-facet Rasch measurement (MFRM) approach to analyze the rating data. This statistical approach has certain advantages over other common approaches, particularly when researchers are working with data obtained from judging plans like the one we employed in this study. Judging plans like ours result in sparse data matrices with much missing data. That is, in our study, the only assessors who rated all the students' essays were the six teachers. The fact that each student only rated

two essays (i.e., the essay that he/she wrote, and one of his/her peer's essays) and not all the essays resulted in an incomplete block design with many "holes" in the resulting data matrix.

Rating data that originate from incomplete block designs can prove to be very challenging for researchers to analyze, particularly if they try to use common classical test theory-based approaches, such as analysis of variance or generalizability theory, given their underlying assumptions. However, the many-facet Rasch measurement approach is well suited to analyzing rating data from incomplete block designs if the judging plan that was implemented results in sufficient linkages. To be sufficiently linked, each element of each facet (i.e., in our study, each assessor, each essay, and each rating criterion) must be able to be linked either directly or indirectly to every other element (Eckes, 2011; Engelhard, 1997; Linacre & Wright, 2002).

Lack of connectedness among the elements of a particular facet, such as the assessor facet, could result in an inability to calibrate all elements of that facet on the same scale (i.e., the researchers would not have succeeded in creating a common frame of reference for interpreting their results). In this type of situation, the researchers would be unable to compare all the measures of the elements of that facet. For example, if the researchers were studying assessor performance, they would be unable to compare all the assessor severity measures because the assessors would not all reside within a common frame of reference. Rather, the researchers would only be able to compare the severity measures of various subsets of assessors (i.e., those particular assessors who had rated students' essays in common).

In our study, the fact that the teachers rated all the students' essays made it possible for us to link all our assessors. Through careful planning, we implemented a judging plan in which every student assessor could be linked to every other student assessor and teacher assessor through the teachers' ratings of the essays. One of the strengths of a many-facet Rasch measurement approach, then, is that in order to make comparisons among assessors, the researchers do not need rating data collected using a complete (fully crossed) rating design. Judging plans that provide that type of data can be very expensive and time consuming to implement. Rather, the researchers need only ensure that in setting up their judging plan, they have built in an appropriate linking network that will guarantee the connectedness of the resulting rating data (for examples of how to set up incomplete but connected judging plans, see Eckes, 2011; Linacre & Wright, 2002).

To analyze the data, we employed Facets, a computer software package (Linacre, 2011). We ran a many-facet Rasch measurement (MFRM) analysis that included assessor type, students' essays, and rating criteria as facets in order to obtain an average severity measure for each of the three assessor types.

A many-facet Rasch measurement model, a logistic latent trait model of probabilities, is an extension of the basic Rasch model that allows researchers to compare not only the difficulty of rating criteria and the proficiencies of students but also the severity of assessors. Linacre (1994) originally developed the MFRM approach to take into account the role that assessors play, acknowledging that not all assessors exercise the same degree of severity when they award ratings. When using a MFRM approach to analyze rating data, assessor consistency rather than exact assessor agreement is required. The Facets computer program simultaneously calibrates all facets included in an analysis and reports the results on an equal-interval logit scale. Thus, Facets reports measures of rating criteria difficulty, student proficiency, and assessor severity using a common frame of reference, which makes it possible to compare elements of the various facets (e.g., individual rating criteria, students, and assessors) both within and across facets, if the data show sufficient fit to the model. A many-facet Rasch model is essentially an additive linear model based on a logistic transformation of observed ratings. The logistic transformation of successive category probabilities function as the dependent variable, and other facets such as student proficiency, rating criteria difficulty, assessor severity and other testing situation facets as independent variables (Engelhard & Myford, 2009). The mathematical formula for the measurement model that we used to conduct our first analysis appears in Appendix C.

We used the results from a fixed-effects chi-square homogeneity test to determine whether there were any statistically significant differences among the three average severity measures for the three assessor types (i.e., self-assessors, peer-assessors, and teacher assessors). The chi-square tested the "fixed-effect" hypothesis that the average severity measures for the three assessor types were the same, after allowing for measurement error. A significant chi-square value would signify that the average severity measures of at least two assessor types were statistically significantly different. We then

performed a multiple comparison procedure (i.e., three unpaired *t*-tests with Bonferroni corrections) to determine whether any of the pairs of average severity measures were statistically significantly different from one another.

Next, we carried out a second MFRM analysis to obtain a severity measure for each individual assessor within an assessor type, as well as an *assessor separation index* for each assessor type. This index is a measure of the number of statistically distinct levels, or strata, of assessor severity within an assessor type. The closer the assessor separation index to zero, the more similar the assessors in their levels of severity. Facets output also provided a measure of the *reliability of the assessor separation*, which indicated how well one could differentiate among the assessors in that group in terms of their levels of severity. It is a measure of how *different* the assessors are. The closer the reliability to zero, the more interchangeable the assessors in terms of severity. The mathematical formula for the measurement model that we used to conduct the second analysis appears in Appendix C.

To run our second analysis we used Facets' grouping capability. Because the student assessors functioned as both self-assessors and as peer-assessors, we identified each of those assessors using two different assessor IDs in the specification file (i.e., we treated each student assessor as if he/she were two assessors) so that we could obtain a measure of the severity that each student exercised when rating his/her own essay, as well as a measure of the severity the student exercised when rating a peer's essay. When we ran this analysis, we fixed the mean of each assessor type at zero but then allowed individual assessors within each type to float relative to that mean.

5.2. Results from our preliminary analysis

Before attempting to answer the research questions we posed, we ran a preliminary Facets analysis to test for data-model fit. The results of the analysis showed that several essays (i.e., essays of students 94, 101, and 160), several assessors (i.e., Assessors 22, 24, 27, 48, 74, 76, 95, 145, and 176), and one rating criterion (i.e., Criterion 7, logical sequencing) were misfitting (i.e., had infit mean-square values equal to or greater than 1.5). Common practice is to delete misfitting elements and then re-run the analysis without them (McNamara, 1996). However, as the purpose of this study was to examine assessor effects, and not to refine a test instrument or assign scores to students' essays, we regarded this approach as inappropriate since unexpected ratings may reveal valuable insights into assessor behavior (Winke, Gass, & Myford, 2011). Therefore, we adopted a different approach.

First, we identified and deleted individual highly unexpected ratings. That is, once we identified misfitting assessors, instead of deleting all the ratings that a misfitting assessor assigned, we instead deleted only those particular ratings the assessor assigned that had standardized residuals equal to or greater than 2. We then re-ran the analysis and this time found no misfitting elements. According to Linacre (2011), satisfactory model fit is indicated when about 5% or fewer of the (absolute) standardized residuals are $\geq$2, and about 1% or fewer of the (absolute) standardized residuals are $\geq$3. Before we began the deletion process, there were 22,113 valid ratings in our data set (i.e., ratings used for estimation of model parameters). Of these, 697 ratings were associated with (absolute) standardized residuals $\geq$2, and 45 ratings were associated with (absolute) standardized residuals $\geq$3, so the number of unexpected ratings was much smaller than the criterion Linacre specified, indicating satisfactory data-model fit.

5.3. Rating scale functioning

The category statistics (Table 1) provide useful information about how well the rating scale performed, indicating to what extent the 6-point rating scale functioned reliably. According to Linacre (2004), in order for a rating scale to perform effectively, a number of guidelines should be met: (a) there should be at least 10 ratings in each category; (b) average measures should advance monotonically with counts; (c) outfit mean-square values should be less than 2; (d) step difficulties (or scale calibrations) should advance monotonically, signifying that each category is the most probable one for assessors to assign to students' essays that are located in a particular portion of the student proficiency continuum; and (e) step difficulties (or scale calibrations) should increase by 1.4, but less than 5 logits. Table 1 shows that the rating scale we used met all these guidelines except for the last one. Fortunately, as Linacre (2004) noted, "this degree of rating scale refinement is usually not required in

Table 1
Category statistics.

Categories	Counts	Average measure	Outfit MnSq	Most probable from
1	875	−.03	1.0	Low
2	1877	.02	.9	−.78
3	3491	.17	1.0	−.51
4	5132	.31	1.0	−.14
5	5317	.44	1.0	.34
6	3007	.60	1.0	1.09

order for valid and inferentially useful measures to be constructed from rating scale observations" (p. 274).

5.4. Variable map

Before we answer our research questions, we first present the variable map we obtained from our analysis in order to familiarize readers with the types of information that one can obtain from a Facets analysis. The map (Fig. 1) displays all facets of the analysis, summarizing key information about each facet.

The first column displays the logit scale, which ranges from 2 to −2 logits. The second column displays measures of the level of proficiency shown in each student's essay, with higher logit measures indicating students whose essays showed higher levels of proficiency. Each star represents seven students' essays, and each dot represents fewer than seven students' essays. The majority of the students' essays were located above the mean. The third column displays the assessor types. Although the severity measures of the three assessor types cluster around the mean, teacher assessors (as a group) tended to rate somewhat more severely on average, while self-assessors (as a group) tended to rate somewhat more leniently on average. Peer-assessors appear midway between those two assessor groups. The fourth column displays the rating criteria. Criteria higher on the scale were harder for students to receive high ratings on, while criteria lower on the scale were easier for students to receive high ratings on. Word choice, sentence variety, and topic relevance were among the most difficult

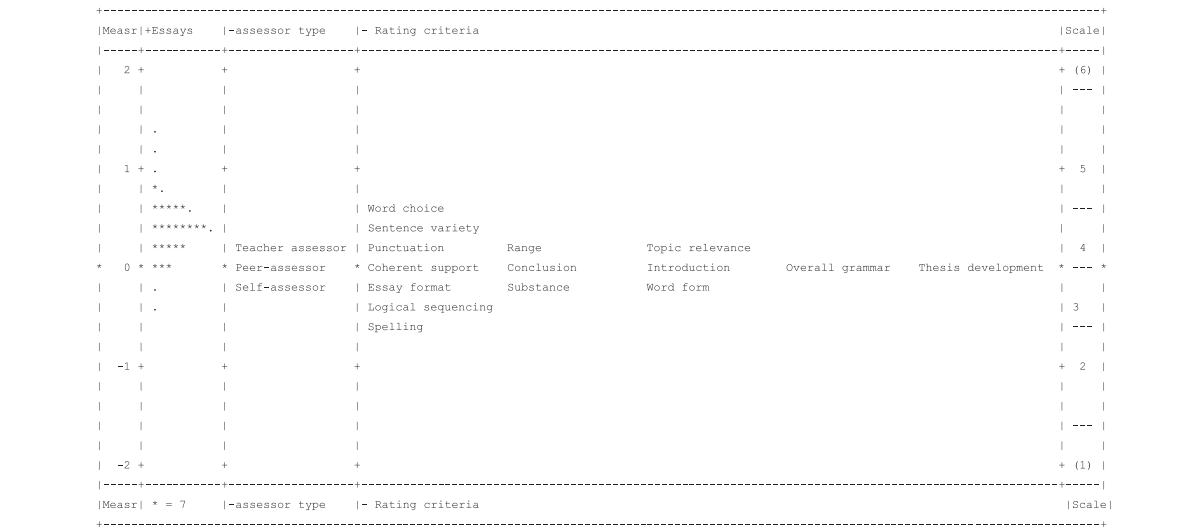

Fig. 1. Variable map showing the ordering of students' essays from those displaying the highest levels of proficiency to those displaying the lowest levels of proficiency, assessor types from most to least severe, and rating criteria from those that were harder for students to receive high ratings on, to those that were easier for students to receive high ratings on.

criteria. By contrast, spelling and logical sequencing were among the easiest criteria. The last column shows the scale structure from category (1) very poor to category (6) excellent.

5.5. Severity differences among assessor types

The average severity measures for the three assessor types, along with their respective standard errors, were as follows: self-assessors (−0.16 logits, 0.02), peer-assessors (0.04 logits, 0.02), and teacher assessors (0.12 logits, 0.01). The results from the chi-square test of homogeneity indicated that the average severity measures for the three assessor types were not all the same, after allowing for measurement error ($\chi^2(2, N = 194) = 205.4$, $p < .0001$).

The average severity measures for the teacher assessors and self-assessors were statistically significantly different, $t(192) = 2.49$, $p = .0135$, 95% CI [0.0586, 0.5014], as were the average severity measures for the peer-assessors and the self-assessors, $t(374) = 7.07$, $p < .0001$, 95% CI [0.1444, 0.2556]. However, the average severity measures for the teacher assessors and the peer-assessors were not statistically significantly different.

5.6. Variability in the levels of severity within each assessor type

While the average severity measures for some of the assessor types differed from one another, there was a great deal of variability among the assessors within each type in terms of the level of severity each exercised. The assessor separation indices and their associated reliabilities suggest that there were multiple statistically distinct levels of severity within each of the assessor types.

The assessor separation index for the self-assessors ($N = 149$) was 3.2, which suggests that there were about three statistically distinct levels of severity within that assessor type. The reliability of the self-assessor separation was .82. Similarly, the assessor separation index for the peer-assessors ($N = 136$) was 3.7, and the reliability of the peer-assessor separation was .86. Finally, the assessor separation index for the teacher assessors ($N = 6$) was 21.53, and the reliability of the teacher-assessor separation was 1.00. Because the teacher assessors evaluated all 188 essays, they assigned many more ratings than did the self-assessors and peer-assessors (i.e., each teacher assigned on average 2570 ratings, while the peer-assessors and self-assessors assigned on average 15 ratings each). Consequently, the teachers' individual severity measures were very precisely estimated, which resulted in an assessor separation index that indicated a greater number of statistically distinct strata than there were actual assessors in that assessor type. The conclusion that is warranted in this situation is that the spread of the teacher assessor severity measures is considerably greater than the precision of those measures (Myford & Wolfe, 2004).

6. Discussion

In the present study, we set out to investigate whether self-assessors, peer-assessors, and teacher assessors differed in the levels of severity they exercised when rating students' essays. We detected and measured severity differences using a many-facet Rasch measurement approach to analyze our rating data. When studying differences in the performance of assessor types, employing a MFRM approach affords certain advantages. As Matsuno (2009) noted, "as more researchers use this research technique, we can illuminate a multitude of facets of self- and peer-assessments" (p. 95). Further, as Basturk (2008) observed, "the facets of the data can be thoroughly investigated individually, which is not always possible in the traditional test analysis" (p. 440).

Compared to previous studies, our study used a rather large number of assessors—188 self-assessors and peer-assessors, and six teacher assessors. Our goal was to compare and contrast the use of self-assessment, peer-assessment, and teacher assessment as approaches for evaluating the quality of students' L2 essay writing. How comparable were the ratings?

When we reviewed the group-level and individual-level statistical indicators, two interesting findings emerged. First, the group-level statistical indicators revealed that the average levels of severity for several assessor types were statistically significantly different. Both teacher assessors and peer-assessors tended to rate significantly more harshly than self-assessors. However, the average levels

of severity of the teacher assessors and the peer-assessors were not significantly different. Second, the individual-level statistical indicators revealed that within each assessor type, assessors exercised a great deal of variability in their levels of severity.

The findings in the present study confirm some of those from previous studies (e.g., we found considerable variability in the severity levels of self-assessors, peer-assessors, and teacher assessors). However, there were some differences, which are understandable, considering the contradictory nature of findings in this area. For example, in Nakamura's (2002) study, teacher assessors were more lenient than peer-assessors. In sharp contrast, Saito and Fujita (2004) and Saito (2008) found that teacher assessors were more severe than peer-assessors. Similarly, Matsuno (2009) reported that peer-assessors tended to rate more leniently than the other assessor types. It is interesting to note that these researchers conducted their studies within the same setting (i.e., Japan), and they all reported variations among assessor types in their levels of severity. Nevertheless, the ordering of assessor types by their average severity levels differed from study to study.

In our study carried out in an Iranian higher education setting, self-assessors were the most lenient of the three assessor types. Our results support those of Sullivan and Hall (1997), who, like us, found that self-assessors tended to overrate themselves. However, our results contradict those of Chen (2008), Matsuno (2009), Brown (2005) and Leach (2000) who indicated that the self-assessors in their studies tended to underrate themselves. The self-assessors in their studies tended to be critical of their writing abilities, the researchers reasoned, and thus assigned lower ratings than they deserved. The researchers hypothesized that students exhibited this tendency to underrate because they did not want to appear boastful and because their cultures valued the importance of modesty. Why would Iranian students tend to assign themselves higher ratings than either peer assessors or teacher assessors? In Iran, students do not share that cultural value of modesty. Student evaluations tend to be norm-referenced. When assigning grades, teachers routinely compare a student's work to other students' work (Farhady & Hedayati, 2009). Consequently, when Iranian students rate their own essays, they are not likely to assign ratings that are lower than those that they would assign to their peers' essays. Because students very much appreciate higher ratings, they may be more likely to assign their own essays higher ratings than they actually deserve.

Our findings support Nakamura's (2002) assertion that one could expect peer-assessors to rate in a reasonably reliable manner alongside their teachers. When we compared the three assessor types, the ratings of the peer-assessors tended to be closer to those of the teacher assessors than to those of the self-assessors (though, on average, the peer ratings tended to be somewhat higher than the teacher ratings). Our results mirror those of Schelfhout, Dochy, and Janssens (2004) who found that the peer assessors in their study were more likely "to overestimate than to underestimate" (p. 194). Possible explanations for this tendency of peer-assessors in our study to overrate stem from cultural beliefs. Iranian peer-assessors may assign higher ratings than peer-assessors from other cultures because Iranian students do not want to be critical of their classmates. Assigning their classmates low ratings may engender animosity, some students believe, thereby ruining friendships and creating the false impression that they are intentionally being critical of their peers. Alternatively, some students may believe that they are showing respect for their classmates when they assign them higher ratings (even though they may realize that their classmates' work does not actually deserve those ratings). A third possible reason relates to Islamic teachings, which stress the importance of thinking of one's neighbors first and one's self second.

Given that there were some differences between the teachers' ratings and the peer-assessors' ratings (though not statistically significant), we would agree with Freeman (1995) who cautioned that "considerable care should be taken in introducing innovative forms of summative assessment which involve elements of subjectivity" (p. 298). Perhaps with some appropriate training, Iranian students could learn to use assessment criteria in a more objective manner to score their peers' essays so that their ratings could be used for summative assessment purposes, but that remains to be seen.

Our finding that self-assessors tended to rate significantly more leniently than the other two assessor types supports Matsuno's (2009) conclusion that "self-assessment was somewhat idiosyncratic and therefore of limited utility as a part of formal assessment" (p. 75). Thus, we would agree with his position that "it is difficult to recommend using self-assessment for formal grading" (p. 950). While this assessment practice may be appropriate for decision making in low-stakes, formative assessment

contexts, our findings would not support its use in high-stakes, summative assessment contexts. Still, is it reasonable to suggest that teachers could work with students to help them learn how to use assessment criteria in a more objective fashion so that over time, and with much guided practice and detailed, explicit feedback on their rating behavior, the ratings they assign their own essays could become part of formal grading procedures?

Like other studies, our results indicate that the brief training we provided did not eliminate differences in the levels of severity that assessors exercised. Some researchers who have investigated the effects of training on subsequent rater behavior have concluded that it is difficult, if not impossible, to change raters' severity levels, even with directive feedback (Knoch, 2011; Lunt et al., 1994; O'Sullivan & Rignall, 2007; Wigglesworth, 1993), though other researchers have reported that rater training can be somewhat effective in reducing differences between raters in the levels of severity they exercise (Elder et al., 2005; Knoch et al., 2007). However, few of these researchers have looked specifically at the effects of training on the rating behavior of students as self-assessors. In a more recent study, Butler and Lee (2010) reported that providing assessor training had only marginal effects in altering the rating behavior of self-assessors. They suggested several reasons why the training seemed to have so little effect. In the paragraphs below, we discuss several of those reasons in the context of our study and then posit some additional ones.

The students in our study did not receive any type of feedback as they carried out the rating task; they participated in a one-hour training session, and then they rated their own essay and a peer's essay. Perhaps receiving feedback might have helped them to use the assessment criteria in a more objective manner.

Students in our study were trained using only a single essay to help them learn to apply the assessment criteria. Perhaps having more than one essay to study and discuss might have helped them to see differences in the quality of students' writing. If they had been trained using multiple essays displaying varying levels of writing proficiency, then that might have helped them to internalize the assessment criteria more fully. Some large-scale assessment programs routinely use a series of benchmark papers to train raters to distinguish between various levels of writing quality. For example, in Iran, benchmark papers of this type are used in training raters to score students' responses to constructed-response items included in the Test of Language, designed and administered by the Iranian Measurement Organization, and to score the three national school-based tests which Iranian students take in the fifth grade of primary school, in the third grade of guidance school, and in the third year of high school. For teachers who participate as raters in these assessment programs, having the opportunity during training to engage in meaningful dialog with their peers as they critically examine and discuss papers of varying quality often functions as valuable professional development. They learn how to use the assessment criteria in a defensible manner to register their judgments. Perhaps if students had the opportunity to engage in similar dialogs with their teachers (and with one another) over student work, they too could learn to function more objectively as self-assessors and peer-assessors.

Students in our study evaluated only two essays. Assessment criteria may need sufficient time to "settle in" so that assessors can learn to use those criteria in an objective fashion. If the self-assessors had had more opportunity to practice scoring sample essays, they may have become more adept at analyzing essays in order to see their strengths, as well as areas in which they needed improvement. The ordering of the essays that students scored may also have played a role. Students evaluated their own essay first, and then a peer's essay. Perhaps if each student had rated a number of their peers' essays before they scored their own, they might have been more objective in assessing their own capabilities.

Students in our study had no hand in creating the assessment instrument. Thus, they had no opportunity to identify and discuss what the criteria ought to be for evaluating essays; rather, they were given an intact assessment instrument to use, the only familiarity with which was through the one-hour training. Perhaps if students had worked in collaboration with their teachers to design the assessment instrument and then had practiced using it to evaluate essays that their peers wrote, they might have obtained a deeper understanding of (and investment in) the assessment criteria and would have been better prepared to assess more objectively the essays we used in our study.

7. Conclusions

It is still premature to suggest that teachers use peer-assessors' ratings when they are making summative judgments about the writing ability of students in their courses, given the hotly debated nature of such summative techniques, and considering the unresolved reliability and validity issues (Ross & Starling, 2008; for more recent discussions, cf. López-Pastor, Fernández-Balboa, Pastor, & Aranda, 2012; Weaver & Esposto, 2012). Researchers have carried out the majority of the studies of peer-assessment in L1 settings, and many of these researchers have studied this assessment practice when teachers were using it for informal, formative purposes. Researchers studying peer-assessment when used for formal, summative purposes have reported mixed results regarding its utility (Cheng & Warren, 1999; Kwan & Leung, 1996). In our study, the average levels of severity of the teacher assessors and the peer-assessors were not significantly different, which might seem to lend at least partial credence to such a role for peer-assessors. However, it is important to point out that we studied students from a single cultural background (i.e., Iranian) who were assessing essays in an L2 setting. We do not know whether our findings are applicable to other settings in which peer-assessors rate the writing performance of their fellow students (or to peer-assessors from other cultural backgrounds). We propose that researchers continue to explore the use of peer-assessment for summative purposes in other disciplines, fields of study, and cultures in EFL settings in order to gain a fuller understanding of its potential.

Results from our study suggest that cultural mores might exert an influence on students' abilities to self assess and peer assess. In the future, researchers might consider carrying out cross-cultural studies of self-assessment and peer-assessment practices to determine to what extent cultural values and expectations might play a role in determining how students perform these rating tasks.

Appendix A.

Essay rating sheet

Essay number: Assessor's name:	Very poor	Poor	Fair	Good	Very good	Excellent
1. Substance	1 □	2 □	3 □	4 □	5 □	6 □
2. Thesis development	1 □	2 □	3 □	4 □	5 □	6 □
3. Topic relevance	1 □	2 □	3 □	4 □	5 □	6 □
4. Introduction	1 □	2 □	3 □	4 □	5 □	6 □
5. Coherent support	1 □	2 □	3 □	4 □	5 □	6 □
6. Conclusion	1 □	2 □	3 □	4 □	5 □	6 □
7. Logical sequencing	1 □	2 □	3 □	4 □	5 □	6 □
8. Range	1 □	2 □	3 □	4 □	5 □	6 □
9. Word choice	1 □	2 □	3 □	4 □	5 □	6 □
10. Word form	1 □	2 □	3 □	4 □	5 □	6 □
11. Sentence variety	1 □	2 □	3 □	4 □	5 □	6 □
12. Overall grammar	1 □	2 □	3 □	4 □	5 □	6 □
13. Spelling	1 □	2 □	3 □	4 □	5 □	6 □
14. Essay format	1 □	2 □	3 □	4 □	5 □	6 □
15. Punctuation/capitalization/handwriting	1 □	2 □	3 □	4 □	5 □	6 □

Appendix B.

Guidelines for how to use the essay rating sheet[1]

1. Having a good command of, and full knowledge about, the topic at hand/including a wide variety of content and providing main ideas to develop the topic
2. Writing a thesis statement and a blueprint
3. Relevance of the essay to its topic and unity of the essay

[1] This is an English translation of these guidelines from Farsi.

4. Having an introduction (including a motivator, a thesis statement, and a blueprint)
5. Having a body and topic sentences (three central paragraphs with a topic sentence for each central paragraph, features of a good topic sentence, unity and coherence of central paragraphs, expressing ideas to develop the topic sentences in each central paragraph)
6. Having a conclusion (including both a reworded thesis statement and a clincher)
7. Appropriate use of supporting ideas in the central paragraphs and placing and developing the blueprint in the respective paragraphs (the first item in the blueprint in the first central paragraph, the next item in the second, and the last item in the third)
8. Use of a wide variety of words/correct use of words/appropriate use of words
9. Appropriate choice and use of idioms, proverbs, and collocations
10. Correct word forms in the sentences
11. A wide variety of sentence structures/use of different types of sentences from simple sentences to complex–compound sentences/effective use of sentence structures
12. Subject–verb agreement/correct use of tenses/correct word order/correct use of definite and indefinite articles/correct use of pronouns and all the grammatical structures at large
13. Correct spelling
14. Essay format (the essay should be five paragraphs long and vary in length between 500 and 700 words.)
15. Appropriate use of punctuation marks/use of small and capital letters when necessary/legible handwriting

Appendix C.

The mathematical formula for the measurement model we used to conduct our first analysis appears below:

$$Log\left(\frac{P_{nirk}}{P_{nir(k-1)}}\right) = B_n - D_i - T_r - F_k$$

where: P_{nirk} is the probability that a student's essay n will receive a rating of k on criterion i from an assessor of assessor type r, $P_{nir(k-1)}$ is the probability that a student's essay n will receive a rating of $k-1$ on criterion i from an assessor of assessor type r, B_n is the level of proficiency shown in student's essay n, D_i is the difficulty of criterion i, T_r is the average severity of assessors in assessor type r, and F_k is the difficulty of scale category k, relative to scale category $k-1$.

The mathematical formula for the measurement model we used to conduct our second analysis was as follows:

$$Log\left(\frac{P_{nirk}}{P_{nir(k-1)}}\right) = B_n - D_i - C_j - F_k$$

where: P_{nijk} is the probability that a student's essay n will receive a rating of k on criterion i from assessor j, $P_{nij(k-1)}$ is the probability that a student's essay n will receive a rating of $k-1$ on criterion i from assessor j, B_n is the level of proficiency shown in student's essay n, D_i is the difficulty of criterion i, C_j is the severity of assessor j, and F_k is the difficulty of scale category k, relative to scale category $k-1$.

References

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education*, *17* (2), 199–214.

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment & Evaluation in Higher Education*, *33* (4), 431–444.

Blanche, P., & Merino, M. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39* (3), 313–340.

Blue, G. (1994). Self-assessment of foreign language skills: Does it work? *CLE Working Papers*, *3*, 18–35. (ERIC Document Reproduction Service, No. ED396569).

Brown, A. (2005). Self-assessment of writing in independent language learning programs: The value of annotated samples. *Assessing Writing*, *10* (3), 174–191.
Brown, H. (2004). *Language assessment: Principles and classroom practices*. New York, NY: Longman.
Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27* (1), 5–31.
Chen, Y. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, *12* (2), 235–262.
Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education*, *24* (3), 301–314.
Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22* (1), 93–121.
Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.
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.
Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–33.
Engelhard, G., & Myford, C. M. (2009). Comparison of single- and double-assessor scoring designs for the assessment of accomplished teaching. *Journal of Applied Measurement*, *10* (1), 52–69.
Evans, A. W., McKenna, C., & Oliver, M. (2005). Trainees' perspectives on the assessment and self-assessment of surgical skills. *Assessment & Evaluation in Higher Education*, *30* (2), 163–174.
Falchikov, N. (1995). Peer feedback marking: Developing peer-assessment. *Innovations in Education & Training International*, *32* (2), 175–187.
Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59* (4), 330–395.
Falchikov, N., & Goldfinch, J. (2000). Student peer assessment: A meta-analysis comparing peer and teacher remarks. *Review of Educational Research*, *70* (3), 287–322.
Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, *29*, 132–141.
Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, *20* (3), 289–300.
Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20* (4), 304–315.
Goldfinch, J. M., & Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment & Evaluation in Higher Education*, *15* (3), 210–225.
Hamayan, E. V. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics*, *15*, 212–226.
Hargreaves, L., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal*, *39* (1), 69–96.
Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
Jacobs, G. M., Curtis, A., Braine, G., & Huang, S.-Y. (1998). Feedback on student writing: Taking the middle path. *Journal of Second Language Writing*, *7* (3), 307–317.
Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
Jiliang, C., & Kun, W. (2007). An investigation of scorer reliability of self and peer assessment of EFL writing among Chinese college students. *CELEA Journal*, *30* (1), 3–11.
Kirby, N. F., & Downs, C. (2008). Self-assessment and the disadvantaged student: Potential for encouraging self-regulated learning? *Assessment & Evaluation in Higher Education*, *32* (4), 194–475.
Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study. *Language Testing*, *28* (2), 179–200.
Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12* (1), 26–43.
Kwan, K., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment & Evaluation in Higher Education*, *21* (3), 205–215.
Leach, L. (2000). *Self-directed learning: Theory and practice* (Unpublished doctoral thesis). University of Technology, Sydney, Australia.
Leach, L. (2012). Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education*, *37* (2), 137–147.
Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
Linacre, J. M. (2004). Optimizing rating scale effectiveness. In: E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 257–578). Maple Grove, MN: JAM Press.
Linacre, J. M. (2011). *FACETS (Version 3.68.1) [Computer software]*. Chicago, IL: MESA Press.
Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3* (4), 484–509.
Lindblom-Ylänne, S., Pihlajamaki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education*, *7* (1), 51–62.
Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11* (3), 279–290.
López-Pastor, V. M., Fernández-Balboa, J., Pastor, M. L. S., & Aranda, A. F. (2012). Students' self-grading, professor's grading and negotiated final grading at three university programmes: Analysis of reliability and grade difference ranges and tendencies. *Assessment & Evaluation in Higher Education*, *37* (4), 453–464. http://dx.doi.org/10.1080/02602938.2010.545868

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12* (1), 16–33.
Lunt, H., Morton, J., & Wigglesworth, G. (1994 July). *Rater behaviour in performance testing: Evaluating the effect of bias feedback.* Paper presented at 19th annual congress of applied linguistics association of Australia, University of Melbourne.
Maclellan, E. (2004). How convincing is alternative assessment? *Assessment & Evaluation in Higher Education*, *29* (3), 311–321.
Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26* (1), 75–100.
McGarrell, H. M. (2010). Native and non-native English speaking student teachers engage in peer feedback. *Canadian Journal of Applied Linguistics*, *13* (1), 71–90.
McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples* (Unpublished master's thesis). Melbourne, Australia: The University of Melbourne.
McNamara, T. F. (1996). *Measuring second language performance*. New York, NY: Longman.
Mok, J. (2011). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal*, *65* (3), 240–250.
Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using manyfacet Rasch measurement: Part II. In: E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460–517). Maple Grove, MN: JAM Press.
Nakamura, Y. (2002). *Teacher assessment and peer assessment in practice (Educational Studies 44)*. Tokyo, Japan: International Christian University. (ERIC Document Reproduction Service No. ED464483).
O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In: L. Taylor & P. Falvey (Eds.), *IELTS collected papers*. Cambridge, UK: Cambridge University Press.
Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, *21* (3), 239–250.
Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, *6* (1), 1–13.
Percival, F., & Ellington, H. (1984). *A handbook of educational technology*. London, UK: Kogan Page.
Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education*, *26* (3), 235–246.
Ross, A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation*, *11* (10), 1–13. Available online: http://pareonline.net/getvn.asp?v=11&n=10
Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15* (1), 1–20.
Ross, S. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics*, *26* (3), 317–342.
Ross, A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of grade 9 geography. *Assessment in Education: Principles, Policy, & Practice*, *15* (2), 183–199.
Sadler, P. M., & Good, E. (2006). The impact of self-and peer-rating on student learning. *Educational Assessment*, *11* (1), 1–31.
Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, *25* (4), 553–581.
Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, *8* (1), 31–54.
Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. *RELC Journal*, *40* (2), 149–171.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25* (4), 465–493.
Schelfhout, W., Dochy, F., & Janssens, S. (2004). The use of self, peer and teacher assessment as a feedback system in a learning environment aimed at fostering skills of cooperation in an entrepreneurial context. *Assessment & Evaluation in Higher Education*, *29* (2), 177–201.
Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: The value added of the students' perspective. *Studies in Higher Education*, *26* (3), 327–343.
Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18* (4), 373–391.
Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76* (1), 7–33.
Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment & Evaluation in Higher Education*, *22* (3), 289–306.
Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68* (3), 249–276.
Topping, K. J. (2003). Self- and peer-assessment in school and university: Reliability, validity and utility. In: M. Segers & E. Cascallar (Eds.), *Optimizing new methods of assessment: In search of qualities and standards* (pp. 55–87). Dordrecht, Netherlands: Kluwer Academic Publishers.
Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48* (1), 20–27.
Topping, K. J. (2010). Peers as a source of formative assessment. In: H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 69–75). New York, NY: Routledge.
van Gennip, N. A. E., Segers, M., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal and structural features. *Learning and Instruction*, *4* (1), 41–54.
Vu, T., & Dall'Alba, G. (2007). Students' experience of peer assessment in a professional course. *Assessment & Evaluation in Higher Education*, *32* (5), 541–556.
Weaver, D., & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment & Evaluation in Higher Education*, *37* (7), 805–816.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11* (2), 197–223.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15* (2), 263–287.
Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10* (3), 305–323.

Winke, P., Gass, S., & Myford, C. M. (2011). *The relationship between raters' prior language study and the evaluation of foreign language speech samples (TOEFL iBT™ Research Report No. 16)*. Princeton, NJ: The Educational Testing Service.

Rajab Esfandiari is an assistant professor at Imam Khomeini International University in Qazvin, Iran. His areas of interest include teaching and assessing L2 writing, many-faceted Rasch measurement, and L2 classroom assessment.

Carol M. Myford is an associate professor of Educational Psychology at the University of Illinois at Chicago. Her areas of specialization include scoring issues in large-scale performance and portfolio assessments, the detection and measurement of rater effects, scoring rubric design, and quality control monitoring of rater performance.