The Journal of Teaching Language Skills (JTLS)
7 (1), Spring 2015, Ser. 78/4
ISSN: 2008-8191. pp. 143-173

Diagnosing the Iranian L2 Writing Ability Using Self-Assessment and Level Specific Approaches

Mahboubeh Taghizadeh*
PhD, TEFL
Iran University of Science and Technology
mah_taghizadeh@ut.ac.ir

Seyyed Mohammad Alavi
Associate Professor, TEFL
University of Tehran
smalavi@ut.ac.ir

Abbas Ali Rezaee
Associate Professor, TEFL
University of Tehran
aarezaee@ut.ac.ir

Abstract

The objectives of this study were (a) to examine the writing performance of L2 learners on the level-specific tasks based on Common European Framework of Reference (CEFR) and (b) to study the likely difference between students' self-assessed level of writing and those reported by raters. The study was conducted with 138 Iranian students at BA and MA levels in Alborz Institute of Higher Education. The participants' majors were Teaching English as a Foreign Language (TEFL), English Literature, and Translation Studies. DIALANG writing self-assessment grid, CEFR writing self-assessment grid, and three writing tasks at B1 (i.e., intermediate), B2 (i.e., upper intermediate), and C1 (i.e., advanced) levels were administered. Descriptive statistics and chi-square test were run to determine each student's writing level based on their performance on the writing tasks and the self-assessment grids. The results showed that (a) no one in the BA group was placed at the C1 level, and only 17.3% of MA students could reach this level; (b) students of both groups rated their writing ability higher on the CEFR grid, whereas they rated themselves lower on the DIALANG grid; and (c) the learners' self-assessment did not correspond closely with their performance on the writing tasks, and only one-third of them were accurate in assessing their writing ability. Writing teachers are suggested to implement self-assessment and introduce CEFR and DIALANG statements as part of the language instruction and to train students to conduct self-assessment based on the *can do* statements.

Keywords: CEFR, DIALANG, diagnostic test, level specific rating, self-assessment, task complexity, writing task

Received: 07/27/2015 Accepted: 12/09/2015

* Corresponding author

1. Introduction

The introduction of alternative assessment in the early 1990s, as Esfandiari and Myford (2013) assert, opened up new opportunities for language classrooms, language education, as well as language assessment. Alternative assessments, as Butler and Lee (2010) state, consist of Self-Assessment (SA), peer-assessments, classroom observations by teachers, student portfolios, and interviews. The focus of this study is on SA which is defined by Andrade, Du, and Mycek (2010) 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). In this research, the writing SA statements of CEFR and DIALANG projects are focused.

The CEFR was developed by an international team of experts working under the aegis of the Language Policy Division of the Council of Europe (Little, 2007), and it has the outcome of 40 years of work on modern languages in various projects of the Council of Europe (Heyworth, 2006). This activity led to a series of detailed syllabus specifications for the different language learning levels, namely the Threshold (i.e., intermediate) (van Ek, 1977), the Waystage (i.e., elementary), and the Vantage (i.e., upper intermediate) Levels (van Ek & Trim, 1991, 1997).

DIALANG is "an on-line learner-centered diagnostic language testing system" (Alderson, 2005, p. 29) and is an assessment system intended for language learners who want to obtain diagnostic information about their proficiency. Ross (1998) notes that the *can-do* type items from the DIALANG may give more accurate results for SA than other formats. Additionally, Brantmeier and Vanderplank (2008) state that for the DIALANG project, Alderson specifically excludes SA from high stakes testing and regards SA more as a valuable descriptive and explanatory tool for providing feedback to learners.

Despite the abundance of research on SA in language learning (e.g., Alderson, & McIntyre, 2006; Chen, 2008; Deakin-Crick, Lawson, Sebba, Harlen., & Yu, 2005; Escribano & McMahon, 2010; Hana Lim, 2007; Kato, 2009; Suzuki, 2009; Wagner & Lilly, 1999), very little empirical examination has been undertaken using CEFR and DIALANG statements in assessing writing ability in foreign language learning context. Many task-based language teaching studies have investigated oral language production and, accordingly, there is a paucity of task-based research on written

language production (Ong & Zhang, 2010). Moreover, in the literature on both L1 and L2 writing, it has been suggested that some task types result in lower test scores than others; however, the relationship between task type or task complexity and writing performance is by no means clear (Kuiken & Vedder, 2008). For instance, studies on the message writing and report writing tasks, which are focused in this study, are scarce despite the fact that these genres are important both in the language teaching pedagogy and in the assessment of foreign language competence.

In this study, a level specific approach was taken. In other words, an approach in which the tasks used were each targeted at one specific level; the written responses of the students were assessed by having trained raters assign a fail/pass rating using level-specific rating instruments. The purposes of this study thus were (a) to determine a writing level based on CEFR for both BA and MA students; and (b) to compare the participants' self-reporting of their writing ability on the CEFR and DIALANG self-assessment statements with their actual writing performance reported by the raters. The research questions formulated in this investigation are as follows:

1. Is there any significant difference in the writing performance of L2 learners on the level-specific tasks based on CEFR?
2. Is there any significant difference between the participants' self-assessed level of writing on the CEFR grid and the writing level reported by the raters?
3. Is there any significant difference between the participants' self-assessed level of writing on the DIALANG grid and the writing level reported by the raters?

2. Review of the Related Literature

2.1 Self-assessment

Moritz (1996) considers SA in foreign language education as a non-traditional form of assessment and a logical component of both a learner-centered pedagogy and self-directed (autonomous) learning programs. Conceptually, SA is supported by theories of cognition, constructivism, and learner autonomy, especially those of Piaget and Vygotsky (Chen, 2008). Deakin-Crick et al. (2005) also suggest that SA builds upon "student self-awareness, student ownership of their own learning process and student responsibility for their own learning" (p. 5). They further argue that "involving learners in the assessment of their own learning relates to theories of learning, the recognition of the importance of motivation for

learning, and the value of non-cognitive outcomes" (p. 1). Brown (2004) also suggests principles of autonomy, intrinsic motivation, and cooperative learning as the theoretical justifications for SA. Similarly, Alderson and McIntyre (2006) note that the implementation of SA for students arises out of the belief in student autonomy as an educational goal.

A great number of advantages are suggested for SA, among which are raising the students' level of awareness of the learning process (Benson, 2001; Blanche & Merino, 1989; Kato, 2009; Nunan, 1988; Oscarson, 1989; Todd, 2002; von Elek, 1985); promotion of learner autonomy (Cram, 1995; Dann, 2002; Kato, 2009; Oscarson, 1989, 1997; Paris & Paris, 2001); increasing students' motivation (Barbera, 2009; Paris & Paris, 2001; Sadler & Good, 2006; Todd, 2002); having a long-lasting effect on students' learning (Oscarson, 1989); setting realistic goals and directing their own learning (Abolfazli Khonbi & Sadeghi, 2012; Blanche & Merino, 1989; Butler & Lee, 2010; Oscarson, 1989); discerning their own individual patterns of strengths and weaknesses (Blue, 1994; Esfandiari & Myford, 2013; Saito & Fujita, 2004); monitoring their progress and reflecting on what they needed to do (Barbera, 2009; Butler & Lee, 2010; Esfandiari & Myford, 2013; Hana Lim, 2007; Harris, 1997; Peden & Carroll, 2008; Sadler & Good, 2006; Sally, 2005); facilitating democratic learning process (Oscarson, 1989; Shohamy, 2001); taking responsibility for their own learning (Barbera, 2009; Esfandiari & Myford, 2013; Paris & Paris, 2001; Peden & Carroll, 2008; Sadler & Good, 2006); and promotion of learning (Black & Wiliam, 1998; Oscarson, 1989).

Concerning SA of writing, Schendel and O'Neill (1999) suggested a number of theoretical and practical advantages: (a) it allows students more control over teacher response to their writing; (b) it can show that students can solve the problems arising in their writing; (c) it can provide students with continued goal setting for a writing course; (d) it supports instruction in revision by asking students to think about how to revise their production to suit a particular purpose and audience; (e) it decreases the negative effects of testing or grading writing; and (f) it helps students to become more metacognitive and self-aware of their own production.

Writing teachers such as Bloom (1997), Elbow (1997), and Yancey (1998) carried out SA into their classes through various methods such as grading contracts with criteria defined by the teacher, though sometimes negotiated with students in the class; portfolios in which students reflect on their writing throughout a semester and discuss their production in light of

the teacher's objectives for the course and their purposes for their own development; reflective writing that assesses papers-in-progress; and oral conferences and written texts in which teacher and student evaluate the written production together or negotiate the course grades.

The relationship between SA and learning and teaching contexts is another issue. Since SA can potentially modify the power relationship between students/teachers, some teachers may find it as a challenge to their authority (Towler & Broadfoot, 1992). In this light, Hamp-Lyons (2007) described two conflicting cultures of assessment: an exam culture and a learning culture. A learning culture's focus is on individual learners' improvement in learning, while an exam culture concentrates on learners' mastery of language proficiency with regard to that of norms or groups. Hamp-Lyons further states that the transition from an exam culture to a learning culture is a complex process, which one requires to take into account the teachers' viewpoints in order to make the transition successful.

The inherent subjectivity of SA as a measurement tool, as Butler and Lee (2010) argue, has traditionally been reported as a threat to its validity. As a result, research analyzing the measurement aspect of SA in foreign and L2 education has predominantly been interested in examining the validity of SA (Butler & Lee, 2010). Butler and Lee further state that "such validation studies have often examined the correlations between self-assessment scores and scores obtained through various types of external measurements such as objective tests, final grades, and teachers' ratings" (p. 7).

In their extensive review of educational SA studies, Falchikov and Boud (1989) and Suzuki (2009) report that more proficient students tended to assess themselves lower compared with teacher marking. Similarly, Blanche and Merino (1989), Blue (1994), and Orsmond, Merry, and Reiling (1997) also found that more proficient students tended to underrate themselves, while less proficient students tended to overrate themselves. However, Topping (2003) notes that (a) scores that students gave to themselves seemed to be higher than scores that teachers gave to them.

Identifying three factors in studies that indicated a close agreement between students' self-assessed ratings and those assigned by teachers, Falchikov and Boud (1989) state that (a) teachers and students assessed more accurately in the better designed investigations, (b) students in advanced courses tended to rate more accurately than those in the introductory courses, and (c) students in science courses rated more accurately than students in other courses. Butler and Lee (2010) have

identified a number of factors for the variability in SA results, which can be broadly classified as "(1) the domain or skill being assessed; (2) students' individual characteristics; and (3) the ways in which questions and items are formulated and delivered" (p. 7).

In addition, some researchers argue that students' cultural backgrounds may impact on how they perform SA practices (Esfandiari & Myford, 2013). For instance, Blue (1994) argues that the students' nationalities and their cultural values might account the discrepancies in SA results with some nationalities tended to overestimate their abilities, whereas others have a tendency to underestimate their language levels. Brown (2005), Chen (2008), and Matsuno (2009), for instance, reported that in Japan, students tend to be critical of their writing abilities and underestimate their writing ability, reflecting the mores such as ego and modesty of their culture.

2.2 Common european framework

As Barenfanger and Tschirner (2008) suggest, the CEFR was developed on the basis of research in second language acquisition, foreign language education, and test research. North (2004) also states that the CEFR draws on theories of communicative competence and language use in order to describe what a language user has to know and do in order to communicate effectively. Considering CEFR as a comprehensive description of language use, Alderson et al. (2009) also argue that the CEFR can be considered, implicitly at least, as a theory of language development.

According to Little (2005), given scales of CEFR, L2 proficiency is defined in terms of three broad bands (i.e., A, B, C), each of which is subdivided offering six levels (A1, A2; B1, B2; C1, C2). Thus, A1 is the lowest level of proficiency defined in the CEFR and C2 is the highest (Alderson et al., 2009). It is argued that CEFR levels were validated in both quantitative and qualitative studies (Alderson, 2002; Hasselgreen, 2003).

CEFR is a framework which has gained momentum as a reference tool for curricula, educational standards, schoolbook publishers and language assessment in Europe and beyond (Harsch & Martin, 2012). It is a prominent example of successful language (education) policy (Council of Europe, 2001; Baker, 2002; Barenfanger & Tschirner, 2008; Morrow, 2004) and is one of the most ambitious examples of the gradual formation, shaping and reshaping, and most recently, implementation of language education policies (Byrnes, 2007). In addition, the CEFR offers a comprehensive and systematic overview of exactly what foreign language learners need to learn

and how they need to learn it (Barenfanger & Tschirner, 2008). Among other things, as Barenfanger and Tschirner (2008) suggest, the CEFR is intended to help language professionals reflect on their current practice and situate and coordinate their efforts.

As Weir (2005) argues, although CEFR provides language educators and practitioners with much useful and valuable information about language proficiency, in its current form it is not "sufficiently comprehensive, coherent or transparent for uncritical use in language testing" (p. 281). Weir identifies four areas of concern with regard to the use of CEFR for the test development: (a) the scales are premised on an incomplete and unevenly applied range of contextual variables/performance conditions (context validity); (b) little account is taken of the nature of cognitive processing at different levels of ability (theory-based validity); (c) activities are seldom related to the quality of actual performance expected to complete them (scoring validity); and (d) the wording for some of the descriptors is not consistent or transparent enough in places for the development of tests.

In a similar vein, Alderson et al. (2009) suggest four practical problems with the use of CEFR scales for test specification: (a) inconsistencies, where a feature might be mentioned at one level but not at another, where the same feature might occur at two different levels, or where at the same level a feature might be described differently in different scales; (b) terminology problems: synonymy or not?; (c) lack of definition, where terms might be given, but are not defined; and (d) gaps, where a concept or feature needed for test specification or construct definition is simply missing. However, as Weir notes, "the CEFR is not seen as a prescriptive device but rather a heuristic, which can be refined and developed by language testers to better meet their needs" (p. 298).

2.3 DIALANG's assessment framework

DIALANG is a large and complex project, which is funded by the European Union (Escribano & McMahon, 2010) and developed by a team of experts at the University of Lancaster (Klimova & Hubackova, 2013). DIALANG, as Alderson (2005) states, contains tests of five language skills or aspects of language knowledge (i.e., Listening, Reading, Writing, Grammar, and Vocabulary) and due to the constraints on the computer-based testing, the CEFR scales for spoken production were ignored (Alderson, 2005).

DIALANG, according to Alderson (2005), is unique in that it attempts the diagnostic assessment of 14 European languages: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish, Irish, Icelandic, and Norwegian. DIALANG's Assessment Framework and the descriptive scales used for reporting the results to the users are directly based on the CEFR (Council of Europe, 2001). This framework, as Haahr and Hansen (2006) assert, summarizes "the relevant content of the CEFR, including the six-point reference scale, communicative tasks and purposes, themes and specific notions, activities, texts and functions" (p. 78). They further state that DIALANG provides learners with various kinds of feedback on the weak and strong points in their language proficiency and constructive advice for further learning.

DIALANG is aimed at adults who want to know their level of language proficiency and who want to receive feedback on the weaknesses and strengths of their proficiency (Council of Europe, 2001). What can be said about the SA instrument that is validated and used for the DIALANG project, however, is that students assess their abilities better at higher levels of language instruction than at lower levels (Alderson, 2005).

SA, as Alderson (2005) notes, is the central component of DIALANG and the SA statements are taken directly from the CEFR, containing a large number of *can do* statements for each skill and at each level (Haahr & Hansen, 2006). However, the wording of CEFR statements was changed from *can do* to *I can* and some statements were also simplified to be used for the certain audience. As Haahr and Hansen (2006) argue, the SA statements in the DIALANG Framework underwent a piloting procedure similar to the test items and the correlation between their calibrated values of difficulty and the original CEFR levels were very high (0,911-0,928). This result indicates that the DIALANG SA statements correspond closely to the original CEFR levels. It was also revealed that the SA statements were equivalent across different languages (Alderson, 2005).

2.4 Task complexity in writing

Task complexity, according to Robinson (2001a, 2001b, 2005, 2007), refers to the cognitive task features which can be manipulated either to decrease or increase cognitive demands placed on the students when they perform a task. Robinson (2001a) suggests three factors for the task complexity including inherent characteristics of the task itself, which is concerned with the nature of input, the task conditions, and the processing operations

required in performing the tasks, and the outcome that is involved. Robinson considers these factors as the heading of task complexity.

In order to suggest the construct of complexity of tasks in writing, it is crucial to consider the demands that writing tasks make at different stages of the writing process (Kormos, 2011). In Kellogg's (1996) model, writing involves three important recursive and interactive processes: formulation, execution, and monitoring. Formulation involves planning the content of the writing and translating ideas into words. In the execution stage, motor movements are used by the writers to create a hand-written or typed text. However, monitoring ensures that the developed text adequately determines the writer's intention, and if mismatches are found, the text is revised. As Kormos (2011) argues, the cognitive complexity of L2 writing tasks can be assumed to be inherent in the formulation stage and can be defined by the demands tasks make on the planning of the content of writing and/or on the linguistic encoding of the content of the text. Complexity inherent in the task itself is generally deemed to derive from the conceptual demands a task required in the planning stage. Considering a threshold level for writing, Schoonen, Snellings, Stevenson, and van Gelderen (2009) argue that "below a certain threshold of FL linguistic knowledge, the writer will be fully absorbed in struggling with the language, inhibiting writing processes such as planning or monitoring" (p. 81). In other words, adequate L2 competence is a significant requirement for the development of higher level writing abilities such as planning, formulation, and revision (Manchon, Roca de Larios, & Murphy, 2009).

Two influential, competing models of task complexity are suggested in the literature: Skehan and Foster's (2001) Limited Attentional Capacity Model and Robinson's (2001a, 2003, 2005) Cognition Hypothesis. These models indicate that how attentional resources are used, coordinated, and directed to different aspects of language production during task completion. Limited Attentional Capacity Model assumes that human has a restricted information processing capacity and the more complicated the tasks are, the more attentional resources the L2 learners require (Skehan, 1998, 2001, 2003). As their attentional limits are reached, students will prioritize processing for meaning over processing the language form (Kuiken & Vedder, 2008). As a result, the accuracy and complexity of the linguistic output will decrease. Unlike Skehan and Foster, Robinson (2001a, 2003, 2005) states that more complicated tasks improve learners' production

complexity and accuracy, and learners can have access to multiple and non-competitive attentional resources.

As Kormos (2011) states, research on the role of task types and the cognitive complexity in second/foreign language writing task has been scarce. For instance, Kuiken and Vedder (2008) note that "in the literature on both L1 and L2 writing, it has been suggested that some task types result in lower test scores than others, but the relationship between task type or task complexity and writing performance is by no means clear" (p. 49). Hamp-Lyons and Mathias (1994) report that in the more difficult tasks judged by the expert raters, students obtained higher scores than in tasks which were considered easier. They suggest that cognitively complex tasks might motivate students to produce better production than cognitively less complex tasks. In another study, Schoonen (2005) found that the type of text students were required to produce in a writing test had a greater impact on scores than the students' writing ability, highlighting the importance of exloring particular characteristics of the task that might result in variance in writing performance.

3. Method

3.1 Participants

This study was conducted with 138 Iranian students at BA (f = 86) and MA (f = 52) levels at Alborz Institute of Higher Education. The participants' majors were Teaching English as a Foreign Language (TEFL), English Literature, and Translation Studies. They were both male and female (20.3% male and 79.9% female) students who ranged in age from 18 to 28. The details of participants' characteristics are presented in Table 1.

Table 1. Frequency and percentage of participants' gender and field of study

Group	Major			Gender		Total
	TEFL	Translation	Literature	Female	Male	
BA	32 (37.2%) 22(25.6%)	32(37.2%)		72(83.7%)	14(16.3%)	86 (62.3%)
MA	25(48.1%)	27(51.9%)	……..	38 (73.1%)	14(26.9%)	52 (37.7%)

In addition, seven expert raters who had MA or Ph.D. in TEFL were asked to rate the participants' productions. They were experienced language

teachers, and some were already experienced test developers. Additionally, a rater trainer, who had a Ph.D. in TEFL and had extensive experience in teaching writing courses at BA and MA levels, was asked to train the raters.

3.2 Instruments and materials

Two SA grids (i.e., DIALANG writing SA and CEFR writing SA) were used in this study (see Appendices A & B). DIALANG grid consisted of 31 writing SA statements, while the CEFR scale consisted of 15 SA statements for the six levels of CEFR (i.e., A1-C2). They were both binary scales with yes/no responses. The detailed information on the different levels of each scale is presented in Table 2.

Table 2. The number of SA statements for the levels of the DIALANG and CEFR scales

Scale	Levels						
	A1	A2	B1	B2	C1	C2	Total
CEFR	2	2	2	3	3	3	15
DIALANG	6	7	7	4	4	3	31

Three writing tasks at three adjacent levels (B1, B2, and C1) were also used in this study. The first task was chosen from the *Real Writing 2* (Palmer, 2008) and asked respondents to write a message to a friend, describing the process of using washing machine. The second task was chosen from *the Real Writing 3* (Gower, 2008), which required the students to write a report on environmental issues. The third task in which the students were asked to write a report on a survey of supermarket customers was selected from the *Real Writing 4* (Haines, 2008).

A rating scale encompassing three levels (B1-C1) with four major criteria (i.e., task fulfillment, organization, vocabulary, and grammar) was the material used in this study (see Appendix C). Each criterion was defined for each of the three proficiency levels by the level-specific statements describing the features and qualities expected at that level. The components of the major criteria were: Task Fulfillment **(**overall written production, overall written interaction**),** Organization **(**coherence and cohesion, thematic development**),** Grammar (accuracy, general linguistic range**),** and Vocabulary (control, range). Moreover, three possible answers for the three selected tasks, which were suggested in *the Real Writing* textbooks, were given to the raters to help them better rate the participants' productions.

3.3 Procedures

The current study was conducted at the beginning of the fall semester in 2013, and the data were collected over a period of three weeks. Initially, the researchers provided the participants with some background about the CEFR and the DIALANG projects and then presented enough information about the goals of the research to persuade them to carefully answer the questionnaires and tasks. This section is divided into three subsections: selection of writing tasks, SA questionnaires, and training the raters.

3.3.1 Selection of writing tasks

To avoid quite a few complications (e.g., some of the respondents only completing one session, the time effects, and the cognitive demands of the writing tasks), instead of administering the tasks in two sessions, the researchers selected fewer tasks and administered them in one session. In other words, from the six levels of the CEFR, A1, A2, and C2 were left out. Another reason for the exclusion of these levels was related to the fact that BA and MA students of English were assumed to have been able to adequately deal with the tasks of A level in that they had already passed some English courses.

The writing tasks were chosen from the *Real Writing (2008)* series, which were written based on the A2, B1, B2, and C1 levels of the CEFR and contained many level-specific authentic writing tasks. The researchers attempted to select the tasks that matched the sample best, and left out the tasks affected by the cultural biases or prior knowledge or the ones which might be perceived as offensive by Iranian students. The validation procedures for the tasks used in this study were based on what the authors of the *Real Writing* textbooks reported for their developed tasks. For instance, they noted they performed a qualitative validation in which a panel of recognized experts was asked to review the developed writing tasks.

After piloting the selected tasks, in order to control the working time and the possible illegal help (e.g., dictionaries, the internet, other persons, etc.), the tasks were administered in the class. The participants were asked to write the three tasks in 75 minutes; that is, 20 minutes for the B1 level task, 25 minutes for the B2 level task, and 30 minutes for the C1 level task.

3.3.2 Self-assessment questionnaires

First, the Persian versions of the DIALANG and CEFR scales were piloted to assess its quality before they were used with the actual participants. In

other words, three Ph.D. holders in TEFL and 28 students assessed the content of the questionnaires. They were asked to check the questionnaires for possible problems and ambiguities. Based on the feedback received from the informants, some necessary changes were made. It is worth noting that for administering the grids, the researchers only gave students the descriptors in the order they appeared in the grids and let them judge for each descriptor whether it applied or not. Then, the revised questionnaires were administered to 138 students in the Alborz Institute of Higher Education and were asked to answer the questionnaires in 45 minutes. In other words, the respondents were asked to read each statement of the CEFR and the DIALANG writing grids and choose 'Yes' if they thought they could do what was described in the statement and 'No' if they could not.

Cronbach's alpha was used to estimate the consistency of the participants' responses to these questionnaires. Reliability coefficients for both grids were as follows: BA respondents (DIALANG: .908; CEFR: .860) and MA respondents (DIALANG: .886; CEFR: .810). Moreover, the coefficients of the writing scales for the both samples were high, indicating that the responses to the items of both scales were acceptable.

3.4 Training the raters

Several procedures were adopted to ensure the reliability and validity of this study; for instance, assessment training was provided to ensure that raters understood what and how to assess. In the introductory session, the raters were provided with an in-depth familiarization with the CEFR and DIALANG projects, test purpose and construct, tasks, rating instruments, and assessment criteria. In addition, ratings were discussed in relation to scripts and their salient features and some practice, both individual and collective were done. Further, one session was devoted to discuss problems and answer questions. After the training session, the writing productions were randomly distributed amongst the raters, and the researchers tried to make sure that no rater assessed a whole booklet. In other words, the three tasks in each booklet were assessed by a different rater. It is also important to note that to implement quality control measures double coding was done and each production was rated by two judges.

The raters were asked to holistically rate the tasks based on the CEFR rating scale and assess to rate each task on the assumption of fail, i.e., not reaching the targeted level of performance and pass, i.e., reaching the targeted level of performance. The trainer requested the raters to budget 10

minutes for assessing the B level productions and 20 minutes for assessing the C1 level production. They were asked to rate the responses at their homes and return all tasks within a few weeks. The inter-rater reliability for the two ratings was assessed performing Kendall's tau_b. Results showed that rater consistency for B1, B2, and C1 tasks were .938, .904, .945, respectively. This indicated that the raters gave quite the same rating as the inter-rater reliability was quite high.

Due to time constraints and the cognitive demands of the writing tasks, a perfect design was used for this study in which both samples were given the same tasks. In other words, participants were required to perform three writing tasks that corresponded to B1, B2, and C1 levels of CEFR, and A1, A2, and C2 levels were left out. The independent variable was writing performance, whereas the dependent variables were the two SA grids. Gender and field of study were also the control variables for this research.

3.5 Statistical analysis

The detailed analyses of the students' performance on the writing tasks formed the basis for a final overall writing score. The researchers also aggregated all the SA scores for both DIALANG and CEFR grids in order to determine one CEFR self-assessed level per person. In other words, each SA statement was scored with 1 and we assumed that, for instance, a B2 person was expected to 'solve' all A1, A2, and B1 statements plus 75% of B2 statements.

To answer the research questions addressed in this study, the following statistical analyses were performed. Descriptive statistics and chi-square test were conducted to determine each student's writing level based on their performance on the three adjacent writing tasks. In addition, descriptive statistics and chi-square analysis were also performed for SA statements of the CEFR and DIALANG scales.

4. Results

In this section the results of the descriptive statistics and the chi-square analysis for writing tasks, SA statements of the CEFR and DIALANG grids are presented. In addition, descriptive statistics and chi-square analysis for the three categories of underrate, overrate, and match on the SA grids are also discussed.

Table 3. BA students' writing levels in terms of rated performance, DIALANG and CEFR grids in %

Major	Test	A1	A2	B1	B2	C1	C2	Chi-Square	*p*
		Below B1							
TEFL	Rated Performance	25		62.5	12.5	.....	….	13.000	.002
	DIALANG SA	18.8	21.9	21.9	18.8	9.4	9.4	3.250	.662
	CEFR SA	18.8	6.3	21.9	28.1	15.1	9.4	6.250	.283
Translation	Rated Performance	34.4		56.3	9.4	….	…..	10.563	.005
	DIALANG SA	28.1	25	25	3.1	6.3	12.5	11.125	.049
	CEFR SA	12.5	6.3	43.8	15.6	12.5	9.4	17.875	.003
Literature	Rated Performance	31.8		45.5	22.7	….	….	1.727	.422
	DIALANG SA	27.3	36.4	13.6	4.5	4.5	13.6	10..727	.057
	CEFR SA	27.3	9.1	27.3	9.1	4.5	22.7	6.909	.227

As Table 3 indicates, there was no one in the BA group who could reach the C1 level. Literature students performed better on the B2 level task than TEFL and Translation students, whereas the weakest group was Translation students as 34.4 % of these students were at below B1 level. Table 3 also shows that the majority of students in all three majors were placed at B1 level as they were found to perform better at B1 task.

The most frequent (21.9%) self-assessed levels by the TEFL students were related to A2 and B1, whereas the least frequent ones were equally concerned with the C1 and C2 levels on the DIALANG grid. On the other hand, TEFL students mostly (28.1%) assessed themselves at B2 and less at A2 on the CEFR grid. In addition, 21.9% of TEFL students rated themselves at B1 on both DIALANG and CEFR grids. Concerning Translation students, they mostly (28.1%) assessed themselves at A1 on the DIALANG, while only 3.1% assessed their writing at B2. On the other hand, regarding CEFR grid, a significant majority (43.8%) of Translation students assessed themselves at B1, whereas only 6.3% assessed themselves at A2. With regard to English Literature students, the most frequent self-assessed level was A2, while the least frequent (4.5%) levels were B1 and C1 on the DIALANG grid. However, they mostly (27.3%) assessed themselves equally at A1 and B1 and less at the C1 on the CEFR grid.

BA students' performance on the writing tasks, DIALANG, and the CEFR grids can be hierarchically ranked as: TEFL (*rated performance*: B1, Below B1, B2; *DIALANG*: B1, A2, A1, B2, C1, C2; *CEFR: B2, B1, A1, C1,*

C2, A2); Translation (***rated performance***: B1, Below B1, B2; ***DIALANG***: A1, B1, A2, C2, C1, B2; ***CEFR***: B1, B2, A1, C1, C2, A2); Literature (***rated performance***: B1, Below B1, B2; ***DIALANG***: A2, A1, B1, C2, B2, C1; ***CEFR***: B1, A1, C2, B2, A2, C1).

Table 4. MA students' writing levels in terms of rated performance, DIALANG and CEFR grids in %

Major	Test	A1	A2	B1	B2	C1	C2	Chi-Square	*p*
		Below B1							
TEFL	Rated Performance	12		48	20	20	…..	7.480	.058
	DIALANG SA	16	20	32	4	16	12	6.440	.266
	CEFR SA	16	4	16	32	12	20	6.440	.266
Translation	Rated Performance	11.1		40.7	33.3	14.8	…..	6.630	.085
	DIALANG SA	18.5	22.2	25.9	14.8	7.4	11.1	3.889	.566
	CEFR SA	14.8	11.1	33.3	22.2	3.7	14.8	8.333	.139

As shown in Table 4, TEFL students performed better than Translation students on the C1 task; however, about half (48%) of the TEFL students were placed at B1. Table 4 also shows that the majority of Translation students fell between B1 and B2 levels. With regard to B2 task, more students from the Translation group were placed at B2 compared with the TEFL students. Table 4 also indicates that approximately the same number of participants in both groups was considered 'Below B1'. Altogether, the majority of both groups were found to be at B1 level.

As indicated in Table 4, about one-third of TEFL students rated themselves at B1, whereas only 4% assessed their writing ability at B2 on the DIALANG grid. On the other hand, they assessed themselves mostly at B2 and less at A2. The most frequent self-assessed level for Translation students was B1 on both DIALANG and CEFR grids, while the least frequent level was C1 on the both grids. MA students' performance on the writing tasks, DIALANG, and the CEFR grids can be hierarchically ranked as: TEFL (***rated performance***: B1, B2, C1, Below B1; ***DIALANG***: B1, A2, A1, C1, C2, B2; ***CEFR***: B2, C2, B1, A1, C1, A2); Translation (***rated***

performance: B1, B2, C1, Below B1; *DIALANG*: B1, A2, A1, B2, C2, C1; *CEFR*: B1, B2, C2, A1, A2, C1).

Table 5. Percentage and Chi-Square analysis of underrate, iverrate, and match categories on DIALANG and CEFR grid

Major	SA Grids	Underrate	Overrate	Match	Chi-Square	*P*
TEFL	DIALANG	25	37.5	37.5	1.000	.607
	CEFR	21.9	56.3	21.9	7.563	.023
Translation	DIALANG	34.4	28.1	37.5	.438	.804
	CEFR	12.5	53.1	34.4	7.938	.019
Literature	DIALANG	40.9	22.7	36.4	1.182	.554
	CEFR	31.8	40.9	27.3	.636	.727

As indicated in Table 5, respondents rated their writing ability higher on the CEFR grid, whereas they rated themselves lower on the DIALANG grid. The highest matches (37.5%) were related to the TEFL and Translation students' ratings on the DIALANG grid, while the highest mismatch (56.3%) was overestimation of the Translation students on the CEFR grid. Table 5 also shows that equal number (37.5%) of TEFL students overrated and matched on the DIALANG, while above half of them (56.3%) tended to overrate on the CEFR grid. About half of the Translation students tended to overestimate on the CEFR, while about the same number tended to underrate and match on the DIALANG. Table 5 also shows that Translation students received the highest match on the DIALANG, while they mostly overrated on the CEFR grid. Regarding Literature students, 40.9% underrated their writing performance on the DIALANG, while the same percentage of them overrated on the CEFR grid.

Table 6. Percentage and Chi-Square analysis of underrate, overrate, and match categories on DIALANG and CEFR grid

Major	SA Grids	Underrate	Overrate	Match	Chi-Square	*p*
TEFL	DIALANG	48	32	20	2.960	.228
	CEFR	28	56	16	6.320	.042
Translation	DIALANG	48.1	29.6	22.2	2.889	.236
	CEFR	29.6	33.3	37	.222	.895

As shown in Table 6, about half of MA students of both TEFL and Translation underrated their writing abilities on the DIALANG grid, while they mostly overrated their writing performance on the CEFR grid. Concerning TEFL students, half of them exaggerated on the CEFR grid, and only 16% of their rating on the CEFR grid matched with their rated performance. Table 6 also indicates that the highest matches (37%) were achieved by the Translation students on the CEFR grid, while the highest mismatch (56%) was overestimation on the part of the TEFL students on the CEFR grid.

5. Discussion

In spite of six-year compulsory English education before tertiary education and passing at least three years for BA students and more than five years for MA students of studying English language courses in the Alborz Institute, the students in this study did not perform well on the writing tasks, and their writing scores were quite low. It is believed that this result is probably due to a few reasons. First, before entering university, their previous courses are mostly reading focused, and no systematic instruction is offered for the writing skill. Moreover, the lack of standards for the writing proficiency and the lack of predetermined, concrete writing outcomes, and the traditional teacher-centered teaching methods can account for the problems in the writing courses offered in this center.

It was found that most students of this research did not have clear and accurate perceptions of their writing ability and their SAs did not correspond closely with their rated performance. This might be due to the fact that in Alborz Institute students are not often asked to assess themselves and their

writing abilities and are not also involved in the SA in their writing classes, and their assessment is mostly summative in which the results are reported as a single score. The results showed that the number of matches on the DIALANG was more than that on the CEFR grid. This may in part be due to the fact that the number of writing statements for each level on the DIALANG grid is more than that on the CEFR. In other words, it can be claimed that DIALANG could better depict the students' writing performance.

The findings of this study are not in line with those of Hamp-Lyons and Mathias (1994) in that students achieved lower scores on the difficult tasks compared with easier ones and cognitively complex tasks did not motivate learners to produce better production. In addition, the findings of this study are in contrast with those of Robinson (2001a, 2003, and 2005) as more demanding tasks did not improve task accuracy and complexity. Further, the results are not consistent with those of Schoonen (2005), who indicates that the type of task produced by learners is influenced by students' grades rather than their writing ability in that students of the current research were more capable to perform 'message writing' task than report writing tasks based on notes or surveys. However, the results are in line with findings of Kuiken and Vedder (2008), who found that some task types lead to lower scores than others. In this study most students were not able to produce acceptable texts for the 'report writing' tasks.

6. Conclusions

This study aimed to determine a writing level based on CEFR for the participants of this study and to compare the self-reporting of their writing ability on the CEFR and DIALANG grids with their actual writing performance. Three writing tasks and the DIALANG and CEFR writing SA grids were administered to BA and MA students of English. The results showed that (a) no one in the BA group was found to be at the C1 level and only 17.3% of MA students could reach the C1 level; (b) students of both groups rated their writing ability higher on the CEFR grid, while they rated themselves lower on the DIALANG grid; (c) The learners' SA did not correspond closely with their performance on the writing tasks and only about one-third of BA students appeared to be accurate in assessing their writing, while slightly below one-third of MA students tended to accurately assess their writing performance; (d) the highest matches for the BA students were related to the TEFL and Translation students' ratings on the

DIALANG grid, while the highest mismatch was overestimation of Translation students on the CEFR grid; (e) the highest matches for MA group were achieved by the Translation students on the CEFR grid, while the highest mismatch was overestimation on the part of the TEFL students on the CEFR grid.

It is believed that there is the lack of standard testing and rating procedures for writing assessment in the Alborz Institute. Therefore, writing teachers are suggested to use CEFR standards for teaching and assessing writing in order to promote the students' awareness of their level of writing proficiency in terms of learning goals and objectives of the CEFR. Teachers are also suggested to implement SA and introduce CEFR and DIALANG statements as part of the language instruction and to train students to conduct SA based on the *can do* statements. Students can also evaluate their progress in the language skills based on the *can do* statements and then can formulate certain goals for their future progress. To improve learners' writing performance, materials developers can also consider writing proficiency standards offered in the CEFR to develop teaching materials and textbooks as it is believed that CEFR provides concrete learning outcome, which is one of the essential issues in the course design.

It is argued that research on complexity in writing tasks can give insight into the nature of processes involved in the writing ability improvement and students' interlanguage development while performing writing task. The findings of the current study suggest that teachers take into account their students' writing ability, their developmental sequence, as well as the cognitive load of the tasks while using tasks in their writing classes.

It is believed that more research is required into the role that SA plays in the improvement of the writing ability. One important step in improving student's writing might be to ask them to self-assess their writing performance. In future study, students can receive training on writing *can do* statements and then be put in charge of rating their own performance. After that, the impact of this training on their writing proficiency can be examined. The relationship between SA in terms of CEFR and DIALANG statements and factors such as personality traits, learning anxiety, locus of control, and the cognitive style merits further inquiry. Future researchers can also use more qualitative and in-depth interview with learners and instructors about L2 learners' writing performance with respect to *can do* statements.

The limitation of this study was concerned with the fact that the researchers were not certain whether or not the participants would hastily or carefully complete the questionnaires and tasks, although the researchers did their best to select the most appropriate time for collecting the data. The delimitations of this investigation were related to the sample of the current study which was small, and factors such as sociocultural background, L2 proficiency level, gender, and age of writers which were not taken into account.

References

Abolfazli Khonbi, Z., & Sadeghi, K. (2012). The effect of assessment type (self vs. peer) on Iranian university EFL students' course achievement. *Procedia - Social and Behavioral Sciences, 70*, 1552 – 1564.

Alderson, J. C. (Ed.), (2002). *Case studies in the use of the Common European Framework*. Strasbourg, France: Council of Europe.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., akala, S., & Tardieu, C. (2009). Analysing tests of reading and listening in relation to the Common European Framework of reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3-30.

Alderson, J. C., & McIntyre, D. (2006). Implementing and evaluating a self-assessment mechanism for the web-based language and style course. *Language and Literature, 15*(3), 291–306.

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

Baker, C. (2002). Council of Europe: Common European framework of reference: Learning, teaching, assessment. *Journal of Sociolinguistics, 6*(3), 467-470.

Barbera, E. (2009). Mutual feedback in e-portfolio assessment: An approach to the net folio system. *British Journal of Educational Technology, 40*(2), 342-357.

Barenfanger, O., & Tschirner, E. (2008). Language educational policy and language learning quality management: The common European framework of reference, *FOREIGN LANGUAGE ANNALS, 41*(1), 81-101.

Benson, P. (2001). *Teaching and researching autonomy in language learning*. London: Longman.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign language skills. *Language Learning, 39*(3), 313–340.

Bloom, L. Z. (1997). Why I (used to) hate to give grades. *College Composition and Communication, 48,* 360-371.

Blue, G. (1994). Self-assessment of foreign language skills: Does it work? *CLE Working Papers*, *3*, 18–35.

Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System, 36*, 456–477.

Brown, A. (2005). Self-assessment of writing in independent language learning programs: The value of annotated samples. *Assessing Writing, 10*, 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.

Byrnes, H. (2007). Perspectives. *The Modern Language Journal, 91*, 641-645.

Chen, Y. M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235–262.

Council of Europe (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cram, B. (1995). Self-assessment: From theory to practice. In: G. Brindley (Ed.), *Language assessment in action: Developing a workshop guide for teachers* (pp.271–305). NCELTR: Sydney, NSW.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. New York: Routledge.

Deakin-Crick, R., Lawson, H., Sebba, J., Harlen, W., & Yu, G. (2005). *Research evidence of the impact on students of self-and peer-assessment*. EPPI-Centre: London.

Elbow, P. (1997). Taking time out from grading and evaluating while working in a conventional system. *Assessing Writing, 4*, 5-28.

Escribano, I. D., & McMahon, J. P. (2010). Self-assessment based on language learning outcomes: A study with first year Engineering students. *Revista Alicantina de Estudios Ingleses, 23*, 133-148.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*, 111–131.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 330–395.

Gower, R. (2008). *Real writing 3*. Cambridge: Cambridge University Press.

Haahr, J. H., & Hansen, M. E. (2006). *Adult skills assessment in Europe: Feasibility study*. Danish Technological Institute.

Haines, S. (2008). *Real writing 4*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2007). Final report of the longitudinal study on the school-based assessment component of the 2007 HKCE English language examination. Report submitted to the Hong Kong Examinations and Assessment Authority, November.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*, 49–68.

Hana Lim, H. (2007). A Study of self- and peer-assessment of learners' oral proficiency. *CamLing*, 169-176.

Harris, M. (1997). Self-assessment of language learning in formal setting. *ELT Journal, 51*(1), 12-20.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing 17*(4), 228–50.

Hasselgreen, A. (2003). *Bergen "Can do" Project*. Strasbourg, France: Council of Europe.

Heyworth, F. (2006). The common European framework. *ELT Journal, 60,* 181-183.

Kato, F. (2009). Student preferences: Goal-setting and self-assessment activities in a tertiary education environment. *Language Teaching Research*, *13*(2), 177–199.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp.57–71). Mahwah, NJ: Lawrence Erlbaum.

Klimova, B. F., & Hubackova, S. (2013). Diagnosing students' language knowledge and skills. *Procedia - Social and Behavioral Sciences, 82*, 436 – 439.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*, 148–161.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48–60.

Little, D. (2005). The common European framework and the European language portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22*, 321-336.

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*, 645-653.

Manchon, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing processes: Implications for theory. In R. M. Manchon (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp.102–129). Clevedon, UK: Multilingual Matters.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100.

Moritz, C. E. B. (1996, March). *Student self-Assessment of language proficiency: Perceptions of self and others.* Paper session presented at the meeting of American Association for Applied Linguistics, Chicago.

Morrow, K. (2004). Background of the CEE. In K. Morrow (Ed.), *Insights from the common European framework* (pp.3-11). Oxford, UK: Oxford University Press.

North, B. (2004). Relating assessments, examinations, and courses to the CEE. In K. Morrow (Ed.), *Insights from the common European framework* (pp.77-90). Oxford: UK: Oxford University Press.

Nunan, D. (1988). *The learner-centered curriculum*. Cambridge: Cambridge University Press.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity In EFL students' argumentative writing. *Journal of Second Language Writing, 19*, 218–233

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: Tutor and students' perception of performance criteria. *Assessment and Evaluation in Higher Education, 22*(4), 357–369.

Oscarson M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1–13.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (pp.175–187). Dordrecht, Netherlands: Kluwer Academic.

Palmer, G. (2008). *Real writing 2*. Cambridge: Cambridge University Press.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychology, 36*(2), 89–101.

Peden, B. F., & Carroll, D. W. (2008). Ways of writing: Linguistic analysis of self-assessment and traditional assignments. *Teaching of Psychology, 35*(4), 313–318.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

Robinson, P. (2001b). Task complexity, cognitive resources and syllabus design: a triadic framework for examining task influences on SLA). In P. Robinson (Ed.), *Cognition and second language instruction* (pp.287–318). Cambridge University Press: Cambridge.

Robinson, P. (2003). The Cognition Hypothesis: Task design, and adult task-based language learning. *Second Language Studies, 21*, 45–105.

Robinson, P. (2005). Cognitive complexity and task sequencing: studies in a componential framework for second language task design. *IRAL, 43*, 1–32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*(3), 193-213.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing, 15*(1), 1–19.

Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54.

Sally, A. (2005). How effective is self-Assessment in writing? In P. Davidson, C. Coombe, & W. Jones (Eds.), *Assessment in the Arab world* (pp. 307-321). United Arab Emirates: TESOL Arabia.

Schendel, P., & O'Neill, P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6*(2), 199-227.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1–30.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchon (Ed.), *Writing in foreign language contexts: Learning, teaching and research* (pp.77–101). Bristol: Multilingual Matters.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp.167–185). Harlow: Pearson Education Longman.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*, 1–14.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.183–205). Cambridge: Cambridge University Press.

Suzuki, M. (2009). The compatibility of L2 learners' assessment of self- and peer revisions of writing with teachers' assessment. *TESOL Quarterly, 43*(1), 137-148.

Todd, R.W. (2002). Using self-assessment for evaluation. *English Teaching Forum, 40*(1), 16–19.

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.

Towler, L., & Broadfoot, P. (1992). Self-assessment in the primary school. *Educational Review, 44*(2), 137-151.

Van Ek, J. A. (1977). *Threshold level for modern language levels in schools*. London: Longman.

Van Ek. J. A., & Trim, J. (1991). *Waystage 1990*. Cambridge: Cambridge University Press.

Van Ek, J. A., & Trim, J. (1997). *Vantage Level*. Strasbourg: Council of Europe.

Von Elek, T. (1985). A test of Swedish as a second language: An experiment in self-Assessment. In Y. P. Lee, A. C. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 47-55). Oxford: Pergamon Press.

Wagner, L, & Lilly, D. H. (1999). Asking the experts: Engaging students in self-assessment and goal setting through the use of portfolios. *Assessment for Effective Intervention, 25*(1), 31-43.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*, 281-300.

Yancey, K. B. (1998). *Reflection in the writing classroom*. UT: Utah State University Press.

Appendix A. DIALANG Writing Self-Assessment Grid

CEF R Level	
A1	I can write simple notes to friends.
A1	I can describe where I live.
A1	I can fill in forms with personal details.
A1	I can write simple isolated phrases and sentences.
A1	I can write a short simple postcard.
A1	I can write short letters and messages with the help of a dictionary.
A2	I can give short, basic descriptions of events and activities.
A2	I can write very simple personal letters expressing thanks and apology.
A2	I can write short, simple notes and messages relating to matters of everyday life.
A2	I can describe plans and arrangements.
A2	I can explain what I like or dislike about something.
A2	I can describe my family, living conditions, schooling, present or most recent job.
A2	I can describe past activities and personal experiences.
B1	I can write very brief reports, which pass on routine factual information and state reasons for actions.
B1	I can write personal letters describing experiences, feelings and events in detail. I can
B1	describe basic details of unpredictable occurrences, e.g., an accident.
B1	I can describe dreams, hopes and ambitions.
B1	I can take messages describing enquiries, problems, etc.
B1	I can describe the plot of a book or film and describe my reactions.
B1	I can briefly give reasons and explanations for opinions, plans and actions.
B1	
B2	I can evaluate different ideas and solutions to a problem.
B2	I can synthesize information and arguments from a number of sources.
B2	I can construct a chain of reasoned argument.
B2	I can speculate about causes, consequences and hypothetical situations.
C1 C1	I can expand and support points of view at some length with subsidiary points, reasons and relevant examples.
C1	I can develop an argument systematically, giving appropriate emphasis to significant
C1	points, and presenting relevant supporting detail.
	I can give clear detailed descriptions of complex subjects.
	I can usually write without consulting a dictionary.
C2 C2	I can provide an appropriate and effective logical structure, which helps the reader to find significant points.
C2	I can produce clear, smoothly flowing, complex reports, articles or essays that present a case, or give critical appreciation of proposals or literary works.
	I can write so well that native speakers need not check my texts.

Appendix B. CEFR Writing Self-Assessment Grid

A1	1.People can write a short simple postcard, for example sending holiday greetings. 2.They can fill in forms with personal details, for example writing their name, nationality and address on a hotel registration form.
A2	3.People can write short, simple notes and messages about everyday matters and everyday needs. 4.They can write a very simple personal letter, for example thanking someone for something.
B1	5.People can write simple texts on topics which are familiar or of personal interest. 6.They can write personal letters describing experiences and impressions.
B2	7.People can write clear detailed texts on a wide range of subjects related to their interests. 8.They can write an essay or report, passing on information and presenting some arguments for or against a particular point of view. 9.They can write letters highlighting the personal significance of events and experiences.
C1	10.At this level, people can write clear and well-structured text and express their points of view at some length. 11.They can write about complex subjects in a letter, an essay or a report, underlining what they think are the most important points. 12.They can write different kinds of texts in an assured and personal style which is appropriate to the reader in mind.
C2	13.People can write clearly and smoothly and in an appropriate style. 14.They can write complex letters, reports or articles in such a way that helps the reader to notice and remember important points. 15.They can write summaries and reviews of professional or literary texts.

Appendix C. Rating Scale Used for This Study

	Task fulfillment	Organization	Grammar	Vocabulary
B1	**Overall written production:** Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. **Overall written interaction:** **a.** Can write personal letters and notes asking for or conveying simple	**Coherence and Cohesion:** Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. **Thematic development**: Can reasonably fluently relate a straightforward narrative or	**Accuracy:** **a**. Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations. **b**. Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express. **General linguistic**	**Control**: Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations. **Range:** Has a sufficient vocabulary to express him/herself with

	Task fulfillment	Organization	Grammar	Vocabulary
	information of immediate relevance, getting across the point he/she feels to be important. b. Can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision.	description as a linear sequence of points.	**range:** a. Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times. **b**. Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events.
B2	**Overall written production:** Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources. **Overall written interaction:** Can express news and views effectively in writing, and relate	**Coherence and Cohesion:** **a**. Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution. **b**. Can use a variety of linking words efficiently to mark clearly the relationships between ideas.	**Accuracy:** **a**. Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. **b**. Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect. **General linguistic**	**Control:** Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication. **Range:** Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent

	Task fulfillment	Organization	Grammar	Vocabulary
	to those of others.	**Thematic development:** Can develop a clear description or narrative, expanding and supporting his/her main points with Relevant supporting detail and examples.	**range:** **a.** Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. **b**. Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	repetition, but lexical gaps can still cause hesitation and circumlocution.
C1	**Overall written production:** Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. **Overall written interaction:** Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.	**Coherence and Cohesion:** Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices. **Thematic development:** Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	**Accuracy:** Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot. **General linguistic range:** Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.	**Control:** Occasional minor slips, but no significant vocabulary errors. **Range:** Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.