Journal of English Education
JEE
Url: http://usnsj.com/index.php/JEE
Email: editor.jee@usnsj.com

Assessing L2 Writing in the Absence of Scoring Procedures: Construction of Rating Scales in a Cypriot-Greek EFL in-Class Context

AUTHORS INFO

Elena T. Kkese
Cyprus University of Technology;
UCLan Cyprus
elenaKkese@hotmail.com

ARTICLE INFO

o-ISSN: 2502-6909
p-ISSN: 2502-9207
Vol. 3, No. 2, December 2018
URL: http://dx.doi.org/10.31327/jee.v3i2.809

Received	:	24 July 2018
Reviewed	:	24 July 2018
Accepted	:	16 August 2018

Suggestion for the Citation and Bibliography
Citation in text:
Kkese (2018) or (Kkese, 2018)
Bibliography:
Kkese, E. T. (2018). Assessing L2 Writing in the Absence of Scoring Procedures: Construction of Rating Scales in a Cypriot Greek EFL in Class Context. *Journal of English Education*, *3*(2), 46-56. doi: http: //dx.doi.org/10.31327/jee.v3i2.809

Abstract

Two central issues in the assessment of direct writing tests, especially for L2 contexts, refer to the development of these tasks and the scoring procedures. These allow making inferences about the specific test takers' writing ability and provide useful diagnostic information about what aspects of writing are mostly important for raters. This study was concerned with constructing specific rating scales based on written samples by Cypriot-Greek students in an EFL classroom context in an effort to examine and determine what aspects of writing are more important in L2 writing and how are these divided up. The examination of these written samples was conducted using two different approaches in an effort to come up with valid and reliable ways to evaluate L2 written samples. The developed rating scales addressed overall writing ability and spelling accuracy. The findings suggest that more emphasis on L2 writing may be given to accuracy rather than communicative effectiveness drawing attention to the need of involving tasks that provide the opportunity to students to reflect on content or topical knowledge.

Keywords: second language writing, spelling accuracy, holistic scales, analytic scales, Cypriot Greek

A. Introduction

1. *Theoretical considerations*

In language assessment, constructs and construct definitions are seen as a way of conceptualising what it should be tested (Fulcher, 2015; Green, 2014; Hulstijn, 2011; Bachman & Palmer, 2010; and Inbar-Lourie, 2008), even though some researchers consider these definitions to be less useful when explaining observed behaviour in assessment contexts (Kane, 2012). Constructs, or the tested ability, are usually related to one or more aspects of language ability and are expressed in rating scales (Fulcher, 2012) aiming at providing guidance to raters during the scoring process. In the Cypriot-Greek context, though, there are no rating scales available for assessing L2 (second language) writing in English and this problem is more observable in the classroom context. Therefore, raters, who in most cases are also the teachers, end up paying attention to different aspects of performance during the scoring process (Hsieh, 2011). Understanding which aspects of performance raters pay attention to is crucial for the design of test tasks as well as the construction of rating scales (Taylor & Galaczi, 2011).

With reference to the test tasks used in EFL (English as a Foreign Language) in-class writing assessment, these mostly involve direct writing tests that have to do with the production of a complete text. These tests are preferred over indirect assessment since good writing tests should test writing (Grabe & Kaplan, 2014) and usually involve formative (ongoing) assessment that is considered to be a more valid method compared to other methods of assessment. This is because students have to produce out-of-class written texts as the outcome of multiple drafting, feedback, and revision.

Assessing direct writing tests, though, involves subjective marking since it may not be as straightforward as making 'right-wrong' decisions (Alderson, Clapman, & Wall, 1995). The task of the raters/teachers is the assessment of students based on how well they complete a given task and for this they must proceed to more complicated judgements. Assessment is usually based on a rating scale exemplifying the criteria that will be used to evaluate writing. However, in the cases in which the rating scale has a variety of interpretations, raters may end up interpreting the criteria differently and in this way rater reliability may be affected.

Consequently, looking at the multifaceted nature of writing and the several components composing a text puts into question what constitutes effective communication since ideas as well as language are important for clear written communication. The aim of this study, therefore, was to explore what comprises effective written communication in an EFL in-class context where no rating scale was available. The major focus was on which aspects of performance raters pay attention to that led to the creation of rating scales for the specific context by using two different scoring procedures while the issue of rater consistency was also considered.

2. *Rating scales*

The importance of well-designed rating scales cannot be overlooked given that they represent the test construct to be measured. For the grading of direct writing tests, analytic and holistic rating scales are usually used in classroom contexts. Analytic scales involve the independent marking of each feature of writing, which can be weighed separately and differently (multiple scores are assigned to each writing product) whereas holistic scales refer to the assignment of a single score that integrates the inherent qualities of writing (Weigle, 2002; and Huot, 1990). Rating scales can further involve primary or multiple trait scoring, however, these are not often used in classroom contexts but they mostly involve large-scale assessment. In these cases, the rater identifies one or more specific features of the written product. As a result, choosing the kind of rating scale to use may depend on the assessment purpose (Weigle, 2002).

With reference to analytic scoring, this is used to provide detailed information about students' performance in different aspects of writing. Each feature of writing is marked separately and this helps raters focus on the marking construct, which is very useful for inexperienced raters. A well-known scoring guide for ESL was developed by Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hugley (1981) that consists of five aspects of writing differently weighted involving content (30 points), language use (25 points), organisation and vocabulary (20 points each), and mechanics (5 points).

On the other hand, holistic scoring involves the assignment of a single score to a piece of writing based on its overall impression without paying too much attention to any particular aspect of the written product. This kind of scoring may lead to more reliable results but the exact constructs that are being assessed are not clear (Weigle, 2002). A well-known example of an ESL holistic scale is the TOEFL writing scoring guide used for the TOEFL Writing Test that consists of descriptors of the syntactic and rhetorical qualities of six levels of writing proficiency.

Several issues emerge from this brief discussion concerning the kind of rating scale that should be used in a classroom context. One issue refers to the cognitive load involved in the different kinds of scoring. Analytic scales are thought to reduce the cognitive load of having to weigh different features (Barkaoui, 2011; and Weigle, 2002). However, other studies have stressed the high cognitive load in analytic scoring since raters have more decisions to make and more scores to decide (Seedhouse, Harris, Naeb, & Ustunel, 2014; and Bejar, 2012). Therefore, further research is needed into which rating scale has a higher cognitive load as well as how the raters' preferences concerning cognitive style may affect the decision process (Baker, 2012).

Another issue involves reliability and specifically which rating scale has higher reliability. Analytic scales, due to the more decisions involved, are thought to reduce reliability concerning the final grade (Bejar, 2012; and Barkaoui, 2011). Other studies, though, emphasised on the higher reliability of analytic scales (Huot, 1990; and Cumming, 1990). Moreover, different raters may value different aspects of writing while they may approach this decision-making task in their own way (Ellis, Johnson, & Papajohn, 2002; and Weigle, 2002). This stresses the need to focus on the quality of the rating scales and specifically on the need for clearly defined criteria and well-articulated levels for each scale or subscale within an analytic scheme (Bachman & Palmer, 2010).

Choosing the kind of rating scale to use can be a complex issue even if the literature provides several examples of rating scales out of which most of them are either analytic or holistic. In the effort to make a decision, an appeal to Bachman & Palmer's framework of test usefulness (1996) can be of great assistance. Based on this framework, choosing the appropriate testing procedures involves finding the best possible combination of the six qualities of test usefulness and deciding which ones are the most relevant in a given situation. These qualities involve reliability, construct validity, practicality, impact, authenticity, and interactiveness. Nonetheless, a number of factors must also be considered when designing the rating scale. These factors involve questions such as:

1) who will use the scale
    a) constructor-oriented scales for the construction of tests;
    b) assessor-oriented scales for the scoring process;
    c) user-oriented scales for the test users;
2) what aspect(s) of writing are most important and how should these be divided up
3) how many points or scoring levels will be used
4) how will scores be reported
    a) combining scores;
    b) separate scores.

Additionally, developing writing descriptors for the several aspects of the scale is also of great importance. An approach used is to develop the descriptors a priori, that is, to define the ability being measured by the writing assessment in advance and describe a number of levels of attainment from none to complete mastery (Bachman & Palmer, 2010). A different approach is to develop the descriptors empirically through examining the actual texts and/or operational rating of writing performances. Choosing between the two approaches depends on the extent to which one believes that the most important aspects of the construct can be measured on such a scale and to factors concerning the purpose of the assessment.

3. Research Context and Objective

This study focused on writing in an EFL in-class context attempting to address what comprises effective written communication and the importance of spelling in relation to writing fluency. The language under investigation was English since it is a compulsory subject from the first grade onward. For this study, a total number of 152 out-of-class written samples produced by Cypriot-Greek students in L2 English were examined. More specifically, two groups participated in this study referring to an elementary and an intermediate group. The written samples were marked by the teacher/rater twice. In the first time, holistic scoring was employed while in the second time, analytic scoring was used. The aim was to compensate for the use of one rater. The aim of the study was to address three over-arching questions. Specifically, the Research Questions were:

(a) What type of rating scale should be used for grading a writing task so that it will provide valid and reliable scores?
(b) What components of writing does the rater pay attention when assessing a written sample?
(c) What is the importance of spelling accuracy?

B. Methodology

Due to the presence of both quantitative and qualitative modes of research, the design of the present study was a mixed-method one. For the first stage, the quantitative approach was employed since the teacher/rater was concerned with assessing the written samples, counting the objective features, missing features and errors in focus. The findings of the categories that emerged served as an index of the salience of the categories (Krippendorf, 2013). However, for the second stage, the qualitative approach was more relevant given that the existing theory and research literature on rating scales is limited, as in the case of the Cypriot-Greek context. Therefore, for this stage, the emphasis was on judging and evaluating students' performances in order to construct the rating scales. In this manner, the construct and the different categories emerged from the data (Galaczi, 2014). Finally, the two rating scales were used for assessing the written samples and, therefore, a more quantitative approach was employed.

1. Participants

Elementary and intermediate students attending an English tutorial centre in Larnaca were the participants of this study. With regard to the elementary group, this involved eight students while the intermediate group involved thirteen students. The students of the two groups were taught by the same teacher for the past two years and were native speakers of Cypriot-Greek except for two bilingual students who had a parent coming from England. Be that as it may, and even though these two students were born in Cyprus and were attending a public secondary school, they were removed from the sample as their performance would presumably have compromised the overall results. Having an English-speaking parent and possibly being exposed to English at home would have a significant impact on the students' performance. Consequently, the intermediate group involved eleven students.

Elementary students: The elementary group involved students between the ages of 12-13 years old. These students were either in the 6th grade of elementary school or in the first grade of gymnasium (high-school) and had attended English classes for four years. Nonetheless, the final exams of the previous levels did not include essay writing. As a result, these students were just learning to encode their ideas into written text and transfer their knowledge of spelling English words across different contexts and specifically from discrete spelling tests that were exposed so far to written essays.

Intermediate students: This group consisted of secondary school students of different ages ranging from 13 to 16 (first grade of gymnasium – first grade of lyceum) as well as two students who had omitted the pre-intermediate level due to their good knowledge of English. The majority of the students had attended ESL classes for six years and over the past three years had been writing essays. Therefore, they were expected to be able to write more effectively and transfer their knowledge of spelling English words across contexts to achieve effective communication, spelling at a higher accuracy level. By this age, students are able to incorporate their knowledge of the English language and grapheme patterns and rules focusing on word meanings.

2. Research instruments

Writing samples: Out-of-class written samples (ongoing formative assessment) were thought as the most appropriate material because these constitute one of the most traditional formats for spelling assessment. In-class written samples (summative assessment) that constitute the second traditional format for spelling assessment were not preferred since are now considered to be of questionable validity (Grabe & Kaplan, 2014). This is because these involve the production of a single written sample based on a relatively simple prompt and over a relatively limited time, not giving the opportunity to students to prepare adequately.

A number of writing samples were continuously collected from both groups throughout the research period. The written samples covered a range of eight topics that were specifically designed for the students' level and were based on the curriculum. More specifically, the elementary students had to write eight written samples between 80-120 words while the intermediate students had to develop eight different topics using 100-150 words. All written samples were argumentative and had a discursive focus since they involved presenting and developing arguments, expressing and supporting opinions and evaluating ideas. Guidance was provided to the context through instructions (contextualised) while there were no right or wrong answers to the prompts.

Rating scales: In the effort to define what constitutes effective written communication, the teacher/rater developed two rating scales. The first rating scale was developed taking into account the several components of writing that characterise a good piece of writing. The second rating scale was specifically created for spelling accuracy. Both rating scales were empirically derived since the description of levels emerged from examination of actual task performances and were developed for the specific classroom context. The aim of these rating scales was to ameliorate several of the reliability and validity problems that rating scales present especially when these are imported from other settings and for this reason the two scales were based on the findings of the study. Additionally, they were developed taking into consideration other rating scales that exist in the literature (i.e., Jacobs et al., 1981).

Consequently, both types of quantitative and qualitative data analyses were conducted in the present study. In order to analyse the written samples, a quantitative method was utilised involving the assessment of the samples based on general impression marking. This technique led to the construction of two rating scales through the use of qualitative method. The quantitative method was particularly relevant since it was used for the second rating of the written samples based on the two developed scales.

C. Findings

1. Constructing the Rating Scales

General impression marking was used for assessing the 152 out-of-class written samples even though this approach is considered as a less reliable predecessor to the holistic way of scoring. This is because written samples are assigned a general score without any explicit criteria. In this context, the teacher/rater read each written sample quickly and based her score on a 'general impression' between 0-20. Based on the findings, it was revealed that several components of writing were taken into consideration in the scoring process. These components were broadly divided into two categories: ideas and language. These categories could lump into a composite score even though performance could be poor on one component. However, if overall performance was thought to be satisfactory, a high grade was awarded as long as errors did not affect comprehension. This scoring process led to the empirical development of two analytic scales that were used for rating the 152 written samples for a second time.

a) Analytic Scale on Overall Writing Ability

For the first rating scale, the emphasis was on overall writing ability. In the scoring process, seven components of writing were taken into consideration. These involved Content, Organisation, Vocabulary, Grammar, Conventions I, Conventions II, and the component of 'Other' that included further features that occurred in actual students' performance but did not belong to the rest of the components. Whereas each written sample was rated on a 10-point scale, each component was weighted differently depending on how important it was for the overall product as determined by the testing context. More specifically, Content, Vocabulary, and Conventions I were weighted with 1.5 points each, Organisation with 2.5 points, Grammar with 2 points while Conventions II and the component of Other with 0.5 points each. There were also explicit descriptors of performance for most of the components of the scale serving as criteria that contributed to the specific component of writing. The rating scale addressing the overall writing ability is tabulated in Table 1.

Table 1. Empirically Derived Analytic Rating Scale for Overall Writing Ability

Content (1.5)

- relevant to assigned topic
- well-focused, clear, interesting
- adequate development
- sufficient length
- completeness
- relevance, credibility and thoroughness of supporting detail and discussion through experience or knowledge of the topic

Organisation (2.5)

- paragraph level (1)
  1) inviting introduction
  2) satisfying conclusion
  3) logical and effective sequencing of ideas (unified paragraph structure)
  4) focused topic sentences
  5) well-pacing
  6) coherence/cohesion (variety and appropriateness of linking devices)
  7) fluent expression of ideas
- sentence level (1.5)
  1) range of sentence structure
  2) length of sentences
  3) word order
  4) completeness
  5) fragments well used
  6) appropriate connectives (between sentences)
  7) range of sentence beginnings
  8) fluid

Grammar (2)

- grammatical accuracy (proportion of accurate sentences and clauses)
- grammatical range (variety of grammatical features tenses, structures, modals, auxiliaries etc.) – use of complex constructions

Vocabulary (1.5)

- lexical range (no repetitions)
- lexical accuracy (meaning)
- lexical relevance (effective word/idiom choice and usage, collocations)
- words convey message in a clear, precise and natural way (lack of translation)
- appropriate word register (voice)

Conventions I (1.5)

- spelling accuracy
  1) seriousness of errors
  2) range of errors

Conventions II (0.5)

- accurate punctuation that guides the reader through the text
- consistent application of capitalisation skills
- neat, legible handwriting (presentation)

Other (0.5)

Concerning Organisation, the explicit descriptors of performance for this component were addressing two different levels referring to paragraph and sentence. These were differently weighted with descriptors for the sentence level being considered more important receiving more points. With reference to Grammar, mistakes involved the use of double subjects or null subjects (due to first language interference), the use of incorrect prepositions (*esp.* where), and the use of no reference. Additional mistakes referred to the lack of articles, the incorrect verb forms, and problems with the middle voice in the first language.

Moving to Conventions I referring to spelling accuracy, written samples were rated on the seriousness and range of errors. Consequently, more points were obtained when a written sample included errors that were not severe or frequent. Finally, in the derived analytic scale, the scale scores were combined for a total score rather than being reported separately. Consequently, after deciding the score for each component of the rating scale, these points were added resulting in a composite score for each written sample since combining scores is a more reliable approach (Huot, 1990; and Cumming, 1990).

b) Analytic Scale on Spelling Accuracy

For the second rating scale, the emphasis was on spelling accuracy. Four differently weighted categories rated between 0-1.5 points made up this scale and were based on the seriousness and range of errors. Explicit descriptors and examples of errors existed for each category while all the examples were taken from the students' written samples. The rating scale addressing spelling accuracy is indicated in Table 2.

Table 2. Empirically Derived Analytic Rating Scale for Spelling Accuracy

1.5 points

- error free or some undetectable errors
    1) correct spelling includes a range of:
        a) homophones (where, wear, ware)
        b) spelling rules (double consonants, *i* before *e*)
        c) irregular or unusual spellings (trudged, brought)
        d) contractions (don't, won't)
        e) word endings (-ion, -ness, -ship)

1 point

- occasional, infrequent errors that do not hinder understanding of text
    1) correct spelling includes:
        a) words involving silent letters (includ*e*, favourit*e*)
        b) use of spelling rules
        c) words with two or more syllables (always, favourite, fortunately)
        d) many common words spelled consistently (because)
        e) unusual spellings (science)
    2) errors:
        a) are consistent
        b) indicate phonetic/visual patterns (beautiful-*beutiful*)
        c) involve representing all sounds (different-*diffrant*)
        d) involve difficult, challenging words (excellent-*exellent*)

0.5 points

- frequent errors that cause the reader of the text to struggle
    1) correct spelling includes:
        a) high-frequency words (can, are, and)
        b) other consistently spelled words (much, back)
    2) errors involve:
        a) missing sounds or letters (television-*tevision*, relatives-*relativs*)
        b) confusion of sounds (think-*thing*, barks-*darks*)
        c) confusion of letter order (first-*first*, friends-*freinds*)
        d) close attempts (because-*becouse*)
        e) word endings (stopped-*stopt*, useful-*usefull*)
        f) inconsistent spellings (other-*other, ather*)
        g) unrecognisable words (eyes-*ese*)
        h) double letter patterns (usually-*usualy*)

0 points

- frequent and severe errors that make reading and understanding difficult
    1) correct spelling includes:
        a) few easy high-frequency words (I, go)
        b) few initial letters (*be*cause)

c) few known letter sounds (*f*rom)

2) errors:

a) aren't closely related to phonetic/visual patterns (papers-*peipres*)

b) involve transfer from L1 (glasses-*yuaya*)

2. Assessing the Effectiveness of the Derived Scales

Once the two rating scales had been prepared, the teacher/rater proceeded with the assessment of the written samples for a second time. The order in which the written samples were marked in the first time was changed and all the written samples on one topic were rated before proceeding to another. Each written sample was measured on the different categories comprising the rating scale on overall writing ability while for Conventions I, the second rating scale on spelling accuracy was particularly relevant. Finally, the separate scores for overall writing ability were combined in order to secure a grand total.

To this end, separate scores addressing each of the different categories of the first scale on overall writing ability were also used and the scores were then converted into percentages for each student. Afterward, all the percentages that applied to one category were totaled for every written sample and then divided by the number of students to achieve an average score (mean) for each of the categories (X=ΣX/N). Based on the findings, Conventions I was not among the categories that received a high score. The following figure indicates more distinctly the percentages for the written samples obtained by the two groups of the participants.

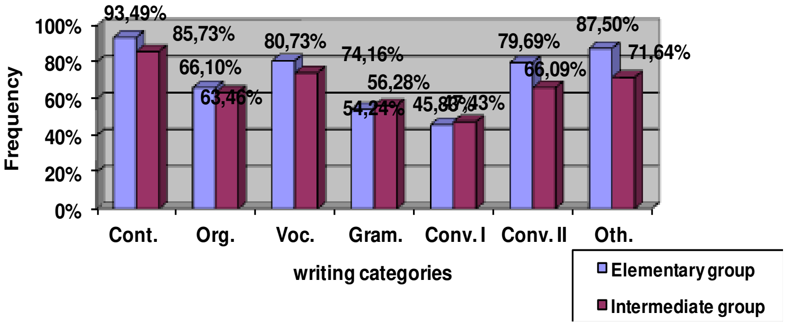

Figure. Average Percentages for the Different Categories Obtained by the Two Groups

The results indicate that the category that received the highest score for the elementary group was Content, followed by Other, Vocabulary, Conventions II, Organisation, Grammar and Conventions I. With reference to the intermediate group, the highest score was also obtained by Content and the lowest by Conventions I but the order of precedence differed for the rest of the categories. Specifically, the second highest score was obtained by the category of the Vocabulary, followed by Other, Organisation, Conventions II, Grammar and Conventions I.

D. Discussion

According to Alderson, Clapham, & Wall (1995), 'the designer of a writing task should also be responsible for designing the scale which will be used to mark the writing' (p. 111). Consequently, the two rating scales that were used for the present study were designed by the teacher/rater. Additionally, guidance for actual scale construction was provided that is not commonly found in the literature. This was achieved by illustrating the procedure for the development of the two scales. Although several textbooks define and illustrate scales, they do not reveal the actual procedures for their development. This lack of direction for testers may account for the current criticism of language rating scales (Weigle, 2002).

The purpose of rating all the writing samples using general impression marking was the careful development of the two empirically derived analytic scales that were used for the second rating. As a result, their development did not begin with a theory of ability levels but description of levels emerged from examination of actual task performances. The derived scales were assessor-oriented aiming at guiding the scoring process and comparing the written samples with the descriptors since assessing subjective marked tests such as written samples can be a difficult task. In this kind of tasks, the rater is required to make complicated judgements since the students' writing cannot be viewed as either correct or incorrect. Consequently, by relying on a rating scale '…is easier for the assessor to decide what level or score to give each learner in a test' (Underhill, 1987, p. 98).

Concerning Research Question (a), analytic scales were preferred over the holistic scales because of the several advantages that the former have and due to the nature of the task in which 'ratings involve subjective judgements in the scoring process…' (Bachman & Palmer, 2010, p. 221). This type of scales '…tend to reflect what raters actually do when rating samples of language use' (Bachman & Palmer, 2010, p. 221) in the cases in which they are based on performance. In the present study, the derived scales are task-specific and address a single population because one rating scale can rarely be used for the assessment of all written or spoken performances (Alderson, Clapham & Wall, 1995). Developing different scales for different tasks or types of tasks is, thus, required. Recognising the difficulty of this process, though, Underhill (1987) suggests 'the only solution is to adapt and improve the [existing] scales by trial and error' (p. 99).

With regard to Research Question (b), the derived scales consist of no more than seven components as it is difficult to make much finer distinctions (Grabe & Kaplan, 2014). Each of the components is weighted differently in proportion to its relative importance to the overall product as determined by the testing context. Moreover, the seventh category (category of 'Other') of the rating scale that addresses overall writing ability includes further features that are not included in the rest of the categories. Explicit descriptors of performance accompany most of the points of the scale since when a scale contains numbers only or descriptors that are simply one-word statements (Excellent, Very Good, Good *etc.*) different examiners may interpret these statements in different ways. The solution, thus, appears to be the development of specific analytic scales including specific descriptors of performance if resources are available, which refers to the practicality of a test (Bachman & Palmer, 2010). Finally, both composite and separate scores were used for assessing writing. More specifically, separate scores indicated the specific categories of writing that the teacher/rater was paying more attention when assessing a written sample.

Regarding Research Question (c), Conventions I that referred to spelling accuracy and the four differently weighted categories that made up this component was one of the two categories receiving the lowest points. On the other hand, Content received more emphasis and, as a result, higher marks than language use for both groups of participants since the only cases in which it could receive a low score involved occasional off-scripts or incomplete responses. The reason was because students could use different writing processes from their first language when writing in the L2. Since the context involved an L2 setting, the most problematic area was thought to be language use. Therefore, the teacher/rater seemed to put more emphasis on linguistic features and generally the linguistic form though which ideas were expressed. This stresses the need to involve more tasks in the EFL classroom that provide students the opportunity to reflect on content or topical knowledge to prepare them better about how to respond to written tasks and about what is needed in such a context (Bøhn, 2015). Separate scores, though, were not used since after deciding the score for each category of the rating scale, the teacher/rater added those points to end up with a composite score for each written sample, since combining scores is more reliable (Huot, 1990; and Cumming, 1990). That composite score was used in decision-making, namely if a certain student did well or not in the specific writing task that was important for his/her overall performance in the course.

E. Conclusion

In the context under investigation, the sample of the written samples was marked twice using a different scoring procedure in each time to compensate for using only one rater. In the first time, the general impression marking scheme was used while in the second time the two analytic scales that were empirically derived were used. In addition, changing the order in which the written samples were marked in the first time and rating all the written samples on one topic before proceeding to another aimed at keeping the standards of grading the same. Finally, the development of the two scales aimed at eliminating the halo effect (Vaughan, 1991) that is commonly observed in classroom contexts. This phenomenon suggests that if a student is considered good/bad in one category, the rater is likely to make a similar evaluation in the other categories as well.

The obtained results are in accordance with the literature suggesting that in L2 writing the emphasis is on language (Bøhn, 2015). The written samples received better results for communicative effectiveness (Content, Organisation, Vocabulary) rather than accuracy (Grammar, Conventions I, Conventions II). This had as an outcome the differentiation of L2 proficiency defined as the control over the linguistic elements from expertise in writing. In L2 contexts, raters evaluate language use more than content and organisation assuming that students can use several of the same writing processes in the L2 as in their first language for content and organisation (transfer of expertise in writing).

Raters, therefore, should focus their efforts upon both the content of ideas and the linguistic form through which these ideas are expressed. They need to coordinate a wide range of complementary concerns when they rate writing samples rather than acting differently in response to different aspects of writing. Using more than one rater would perhaps provide interesting results and ensure rater reliability as long as the raters would not reveal the scores given. However, using more than one rater may entail some disagreement among them since it is not always easy to agree on the exact scores. Similar training is, as a result, required in order to remove idiosyncratic ways of examining that individual raters can easily develop and helps reducing the subjectivity of the raters' decision. Even in such cases, though, "despite their similar training, different markers focus on different essay elements and perhaps have individual approaches to reading essays" (Vaughan, 1991).

F. References

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. United Kingdom, UK: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Baker, B. A. (2012). Individual differences in rater decision-making style: an exploratory mixed methods study. *Language Assessment Quarterly, 9*(3), 225–248. doi: https://doi.org/10.1080/15434303.2011.637262

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

Bejar, I. I. (2012). Rater cognition: implications for validity. *Education Measurement: Issues and Practice*, *31*(3), 2–9. doi: https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bøhn, H. (2015). Assessing spoken EFL without a common rating scale: Norwegian EFL teachers' conceptions of constructs. *Sage Open* (October–December 2015): 1–12. doi: https://doi.org/10.1177/2158244015621956

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. doi: https://doi.org/10.1177/026553229000700104

Ellis, R., Johnson, K. E., & Papajohn, D. (2002). Concept Mapping for Rater Training. *TESOL Quarterly*, *36*(2), 219–233. doi: https://doi.org/10.2307/3588333

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London, UK: Routledge. doi 10.4324/9781315695518

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing*. Oxford, UK: Routledge. https://www.routledgehandbooks.com/doi/10.4324/9780203181287

Galaczi, E. (2014). Content analysis. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 3). Chichester, UK: Wiley-Blackwell. doi: 10.1002/9781118411360

Grabe, W. & Kaplan, R. B. (2014). *Theory & Practice of Writing: An Applied Linguistic Perspective*. New York, NY: Routledge.

Green, A. (2014). *Exploring Language Assessment and Testing: Language in Action*. New York, NY: Routledge. doi: 10.4324/9781315889627

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

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, 229-249. doi:10.1080/15434303.2011.565844

Huot, B. (1990). Reliability, validity, and holistic scoring: what we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*, 385-402. doi:10.1177/0265532208090158

Jacobs, H., Zinkgraf, S. Wormuth, D. Hartfiel, V., & Hugley, J. (1981). *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives*, *10*, 66-70. doi:10.1080/15366367.2012.681977

Krippendorf, K. (2013). *Content Analysis* (3rd ed.). Thousand Oaks, CA: SAGE.

Seedhouse, P., Harris, A., Naeb, R., & Ustunel, E. (2014). The relationship between speaking features and band descriptors: A mixed methods study. *IELTS Research Reports Online Series*. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Taylor, L. & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking* (Vol. 30, pp. 171-233). Cambridge, UK: Cambridge University Press.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques.* Cambridge: Cambridge University Press. doi: 10.1017/S0272263100007361

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9780511732997