

## ASSESSING AND SCORING FOREIGN LANGUAGE LEARNERS' COMMUNICATIVE COMPETENCE

Abeer Alshwiah\*

Ms. lecturer, University of Dammam, Saudi Arabia, email [aalshwiah@uod.edu.sa](mailto:aalshwiah@uod.edu.sa)

### Abstract

Communicative competence plays a vital role in the effective and appropriate communication of foreign language learners (FLLs). Nevertheless, scrutiny of the English language curriculum in Saudi Arabia as regards the development of FLLs' communicative competence will reveal a shortage of valid and reliable test and scoring scales designed to measure this development. Therefore, this study aims to provide a valid and reliable written test for measuring the development of some of the components of FLLs' communicative competence. In addition, it compares two scoring scales, in order to reach a conclusion about some of the characteristics required in a scoring scale for communicative competence. A test was therefore administered to 49 Saudi students at the University of Dammam and scored by three raters using the same scoring scales. The study found that both Holistic and 'Correct Sentences' Scales had a high level of inter-reliability, although neither produced variations in all questions. Ultimately, the Correct Sentences Scale was chosen to score the test, because it categorised the test questions into suitable components. Another important finding was that the test questions measuring the learners' grammatical and discourse competence were reliable and valid, while the questions measuring the learners' sociolinguistic competence were valid, but had low reliability. Therefore, to conclude, the study recommends amending the Correct Sentences Scale, by being more specific about scoring the components in each response. The scale would then be suitable for scoring tests of communicative competence, as well as redefining the test instructions for greater clarity and precision.

**Keywords:** Communicative competence, validity, reliability, assessment, inter-rater.

### 1 INTRODUCTION

Teaching English as a foreign language in Saudi Arabia aims to enable the learners to be competent communicators. Based on Bissenbayeva et al.'s (2013) claim that most foreign language learners (FLLs) do not possess the important skill of communicative competence, we need valid and reliable test and scoring scales to identify their level of communicative competence, particularly within the curriculum offered at the University of Dammam (UD) in Saudi Arabia. Communicative competence, as defined by Hymes (1962), is the speaker's knowledge of grammar, syntax, morphology and phonology, as well as their knowledge of how and when to use the respective language appropriately in a social context. Hymes' (1962) definition of communicative competence awards more attention to the appropriateness of utterance in a social context and emphasizes that it is an important aspect of the effectiveness of communication, in contrast to earlier definitions, such as Chomsky's (1988). Chomsky was purely concerned with ideal speaker-listeners, who

were grammatically competent (having intact syntax and semantics), but lacked pragmatic competence (the ability to use the language appropriately) (Lukin, 2013; McNamara, 1996).

Communicative competence, in this study, refers to the knowledge and skills that enable FLLs to communicate effectively by expressing themselves in coherent text or discourse, appropriate for the context. This definition is based on Canale and Swain's (1980) model of communicative competence, which classifies its components. In this study, due to the limited amount of time allocated to it and the course aims, only grammatical, sociolinguistic and discourse competence will be considered. This is supported by Williams (1979) and Pillar (2012), who assert that it is not compulsory for an English language course to teach the four communicative competences.

## 1.1 The Research Problem

This study aims to provide a valid and reliable written test and scoring scale for communicative competence. A valid test in this study signifies a test which provides trustworthy results and an acceptable interpretation (Chapelle, 1999; Fulcher & Davidson, 2007; Kenyon & McGregor, 2013). On the other hand, a reliable test indicates a test that enables learners to produce the same or similar results if the test is repeated, due to consistent scoring (Carr, 2011; Jones, 2013). To date, there are few studies with this aim. Ayyanathan, Kannammal and Rekha (2012) have measured the development of FLLs' communicative competence in India; Fischer (1984) has measured the development of French language learners' communicative competence and Van Schaik et al. (2014) have measured the development of language learners' communicative competence in the United States. However, there is a dearth of studies that have measured the development of FLLs' communicative competence for this purpose, especially in relation to Foundation Course students at UD in Saudi Arabia. Moreover, there is a dearth of studies that provide a reliable scoring scale for such tests, as Ghanbari, Barati and Moinzadeh (2012) confirm. Due to the fact that it is difficult and critical to decide on a suitable scale, as Barkaoui (2007) confirms, and its reliability represents part of the test's reliability (Henning, 1987; Wang, 2009), this study compares the Holistic Scale adapted by many researchers and a newly adapted scale set to measure development in FLLs' communicative competence. The findings should make an important contribution to the field of testing and assessing communicative competence.

## 1.2 The Research Questions

This research seeks to address the following questions:

1. To what extent is an empirically developed rating scale of communicative competence with level descriptors more valid and useful for assessing FLLs' communicative competence than a Holistic Scoring Scale?
2. To what extent are the test questions valid and reliable?

To answer these questions, the study will test the following hypotheses:

1. The raters will be consistent in rating the students' answers, using the Holistic Scale.
2. The raters will be consistent in rating the students' answers, using the Correct Sentences Scale.
3. The test questions assessing communicative competence are valid.
4. The test questions assessing communicative competence are reliable.

## 2 A REVIEW OF RELATED LITERATURE

### 2.1 Communicative Competence Models

In the literature, there are many models for communicative competence that define its sub-competences and the relationship between them. Hymes' (1972) model was one of the early prominent models of communicative competence and many linguists later improved on its content and features, such as Canale and Swain (1980); Bachman (1990); and Celce-Murcia et al. (1995). Due to the limited scope of this study, only Canale and Swain's (1980) model will be explained in detail, because it is the one that was adopted in this instance. Canale and Swain's model (1980) was reported as the first model to define communicative competence in terms of L2 teaching (Yano, 2002). It defines communicative competence as knowledge of grammatical rules, of how language is used in context to perform functions, and of how utterance and function can be combined, according to discourse principles (Canale & Swain, 1980). The competences in this model are grammatical competence, sociolinguistic competence, discourse competence and strategic competence. This model provides a framework for language testing (Weir, 1990), but it is criticized for its separation of discourse and sociolinguistic competence, because it is claimed that the connectivity of

discourse or a text includes its appropriateness (Schachter, 1996). Moreover, it is criticized for not generating specifications which relate directly to the communicative competence model (Celce-Murcia et al., 1995). Nevertheless, despite such criticisms of Canale and Swain's model, as mentioned earlier, it is still adopted for developing and assessing students' competences, because within it, all competences are regarded as having the same importance and are crucial to successful communication.

### 2.1.1 Assessing Communicative Competence

Different authors have measured the development of learners' communicative competence in a variety of ways. Each has its advantages and drawbacks. Many studies have used pre-test and post-test measures of development before and after the study intervention, such as testing sociolinguistic development, as in Ishihara (2007); Alco'n Soler(2005); Bardovi-Harlig and Vellenga (2012); Strakšienė and Baziukaitė (2010); Ayyanathan et al. (2012), and Agbatogun (2014). Written tests are not time-consuming to conduct, but they need a reliable and suitable scoring scale. Moreover, they do not test oral competence, which is essential in communication. Other studies use oral tests in addition to a written test, such as one conducted by Fischer (1984). Oral tests are more suitable, but also need a reliable scoring scale and are time-consuming to conduct, especially in classes with large numbers of students. On the other hand, Brooks (1992) uses ethnographic observations of 12 students and Bissenbayeva et al. (2013) examined development in the communicative competence of 20 learners through observation, progress analysis, testing and interviewing. Previous studies testing FLLs' communicative competence have dealt with small numbers of participants and this might be the reason for gathering data from multiple sources at various points in time. Due to the large number of students in foundation classes at most universities (about 40 students in each class), it is more convenient to measure development in communicative competence using written tests.

### 2.1.2 Assessing Constructed Responses

The obvious thread running through the assessment of constructed responses, as this study tests, is the raters' variation and subjectivity in assessment. This variation can be due to the raters' different decisions, relating to their background, work experience, bias, characteristics and training in the use of the scale (Fulcher, 2013; Knoch, 2009b; McNamara, 1996). However, there are different approaches that can be used to reduce this variation and achieve consistent scoring. The first approach involves the use of a scoring scale, or what is also referred to as a 'scoring rubric'. There are four types of scoring scales: primary trait scales, holistic scales, analytical scales, and multiple-trait scales (Weigle, 2002). The second approach involves counting the number of errors in a response and dividing the number by the total number of words (Carr, 2011). This approach is very close to the 'objective measures' defined by Perkins (1983), as a means of counting the number of words, clauses or sentences per test unit. This method is considered as an objective and reliable measurement, although it does not consider whether the tense used is the one required in a question, or if the answer is appropriate to the question (Perkins, 1983).

## 3. METHODOLOGY:

### 3.1 Description of the Sample

The test was conducted on an intact sample of 49 English FLLs on the Foundation Course in the College of Applied Science and Community Service at UD in Saudi Arabia.

### 3.2 Description of the Instruments

#### 3.2.1 The Test

The aim in designing the test in this study is to help researchers and teachers ascertain the test takers' communicative competence levels. This test incorporates the features stated by many researchers as being essential in communicative competence tests. These features include: (1) the evaluation of a representative sample of performance (Weir, 1990), (2) the requirement for integrative performance, where the learners show their knowledge suitable for the context (McNamara, 1996; Weir, 1990), (3) like other language tests, a clear definition of the abilities and constructs we are seeking to develop and measure, because these are essential to the test's validity and interpretation of the results (Bachman, 1990; Jones, 2013; Young, 2013), and (4) an assessment of each construct on the basis of at least one task and several integrated tasks, which assess more than one construct (see Table 1 illustrating the definition of the test's constructs and the number of questions in each (Carr, 2011).

**Table 1. The test's constructs**

The construct	Definition	Sub-constructs	Number of questions
Grammatical competence	The learner's ability to use grammatical rules accurately in relation to their needs.	Past tense	4
		Asking wh-questions	3
Sociolinguistic competence	The learner's ability to express accurate meaning using specific functions.	Introducing themselves	2
		Complaining	2
		Responding to a complaint	1
Discourse competence	The ability to produce coherent text.	Using pronouns	4
		Using conjunctions	4

This test is criterion-referenced because the test taker's results are interpreted against a pre-determined, fixed criterion (Carr, 2011). The questions in this test which measure development in the learners' sociolinguistic competence is of a completion task type for written discourse, corresponding to Brown's (2001) classification of pragmatic test types. The reasons for choosing this type of task are; (1) they suit the test's aim of testing the learners' communicative competence in written discourse, where the students are required to read a description of a situation and then write what they would say in such a situation, and (2) its reliability, as Brown (2001) mentions, is very high and validity is moderate. However, this type of testing is challenging, because, (1) it requires multi-raters to rate it using the same scoring rubric (Carr, 2011), and (2) an individual's performance in a test does not indicate that the person will perform the same in other practices, because it measures a sample of the learners' performance (Young, 2013).

### 3.2.2 The Scoring Scales

To assess the test, there was a need to develop a suitable scoring scale, since no suitable ones were found in the literature. For developing the new scoring scales, the following steps, suggested by Weigle (2002) and Plakans (2013), were followed:

1. Identify the type of rating scale (these consist of holistic and correct answer scales),
2. Define the construct that will be tested (see Table 1 above),
3. Identify the scale's users: here, these are FL teachers,
4. Identify the criteria used in assessment (as described in Tables 2 and 3 in Appendix 1),
5. Identify the descriptors (as described in Tables 2 and 3 in Appendix 2), and
6. Determine the way the results will be reported (as described in Tables 2 and 3 in Appendix 1).

A considerable amount of literature has been published on scoring scales. With regard to this study, many scoring scales were found, constructed to assess FLLs' writing in general, such as Bachman and Palmer's (1996), Knoch's (2009b), and Weir's (1990). The scope of these scales consists of assessing different parts of the writing, such as the organization of the text, ideas, content, and other aspects beyond the focus of this research (Brown & Bailey, 1984). However, these scales do not offer an adequate foundation for testing communicative competence, because they do not offer an operationalized definition of assessing such competence (Knoch, 2011).

A number of researchers are more specific, basing their rating scales on Canale and Swain's (1982) communicative competence model, which is the basis of this study; for example, Connor and Mbaye (2002), East (2009), and Hawkey and Barker (2004). It seems that these scoring scales do not take account of communicative competence, but rather concentrate more on scoring levels of proficiency, because they score many aspects. In this study, the researcher has aimed to develop the learners' communicative

competence in just a few aspects, due to the limited time and scope of the study. Therefore, there was a need to develop a new scoring scale. However, developing an efficient and usable scoring scale is a difficult task and even using ready-made scales still needs guidance and practice to check their practicality (Meier, Rich, & Cady, 2006). Moreover, it is difficult to operationalize the scale's qualities to a meaningful performance score (Plakans, 2013).

In this study, a Holistic Scale and an adapted version of the second scoring approach mentioned by Carr (2011) were developed. The holistic score assigns a single score for students' responses, based on the rater's quick reading and on the overall quality or overall level of competence demonstrated (Carr, 2011; Knoch, 2009a; Weigle, 2002). The holistic scoring approach in this study has 5 levels of scoring, ranging from 4 to 0 (See Table 1 in Appendix. 2, 'The Holistic Scoring Scale'). This scale is developed on the basis of three Holistic Scales (Bachman & Palmer, 1996; Knoch, 2009a; Weir, 1990). The advantages of this scale are that it requires 'fewer decisions', and it is time- and cost-efficient (Carr, 2011; Knoch, 2009a). Perhaps the most serious disadvantages of this type of scoring scale are: (1) it does not provide useful diagnostic information about the test taker's competence, because it overlooks some aspects of writing, (2) it does not allow the raters to determine the test taker's level in various areas, such as vocabulary, organization or syntax, and (3) it shows the learners' strengths more than their weaknesses. For these drawbacks, it is compared with another scoring approach.

The second scoring scale was developed by adapting Carr's (2011) second approach to scoring. This approach is referred to here as the 'Correct Sentences' Scale (See Table 2, Appendix. 2, 'The Correct Sentences Scale'). In Carr's (2011) approach, the rater counts the number of errors in a response and divides the number by the total number of words or sentences. This approach is suitable for scoring responses that do not indicate a definite number of sentences in the answer, but rather depend on the learner's ability to answer the questions and convey the meaning. In this research, in order to avoid the disadvantages of this approach - as Carr (2011) states - no credit is given for words which are correct; the raters were merely asked to count the number of sentences they considered to be right, and the number of sentences they considered to be wrong. The raters' scores were then analysed by dividing the number of correct sentences in the task by the total number of correct and incorrect sentences, multiplied by 5. This ensured that all the tasks had a score out of 5 for all the tasks. For scoring wh-questions, a score of 5 was given for each component in the question (see Table 2, Appendix 2).

## 4. THE RESULTS

### 4.1 Scoring Scale Inter-reliability

To test the first two hypotheses: (1) *'The raters will be consistent in rating the students' answers using the holistic scale'*, and (2) *'The raters will be consistent in rating the students' answers using the correct sentences scale'*, the inter-rater reliability was computed. This was carried out by comparing consistency in the ratings of the students' scores, rather than by checking the raters' agreement, which is the approach used with a nominal scale (Vanbelle & Albert, 2009; Wang, 2009). Different statistical methods have been used in the literature to compute inter-rater reliability, namely, (1) correlation coefficients (Alderson, Clapham, & Wall, 1995; Hayen, Dennis, & Finch, 2007; Rezaei & Lovorn, 2010; Sasaki & Hirose, 1999); (2) one-way analysis of variance (ANOVA) (Weigle, 2002); (3) intra-class correlation coefficients (Hayen et al., 2007); (4) Fisher Z Transformation; and (5) the Spearman-Brown Prophecy Formula (Henning, 1987).

The test was rated by two language teachers, in addition to the researcher, as recommended by Green (2014) and using the Holistic and Correct Sentences Scales. To test the first hypothesis, ANOVA was applied to test the raters' variances in using the Holistic Scale for all the test questions. ANOVA was used because it enables multiple comparisons to be made and because the rating was carried out by the raters on an individual basis, not in groups (Hirdes et al., 2002; van Loo, Moseley, Bosman, de bie, & Hassett, 2003). All the items were insignificant at .001. Thus, the first hypothesis is accepted. There was no variation between the raters' scorings using the Holistic Scale. To test the second hypothesis, ANOVA was applied to test the raters' variance in using the Correct Sentences Scale for all the test questions. All the items were insignificant at .001. Thus, the second hypothesis is accepted. There was no variation between the raters' scorings using the scale and so it can be concluded that raters were consistent in scoring the students' responses using it and there was inter-rater reliability in the use of the Holistic Scale and Correct Sentences Scale.

However, when comparing the scales in terms of variation in students' responses (by comparing the means of the scores resulting from the two scales), it is clear that the Correct Sentences Scale shows a variation in students' answers from those questions related to grammatical competence (writing passages in the past), which exceeds what was revealed using the Holistic Scale. The opposite is true in the

sociolinguistic competence questions, where the Holistic Scale shows more variation in students' answers than is the case with the Correct Sentences Scale.

## 4.2 Test Validity and Reliability

To test the third hypothesis '*The test questions assessing communicative competence are valid*' and since the three raters were consistent in their scoring, the averages of the 49 students' scores were used to compute the test's validity and reliability. The test validity was measured using factor analysis for the Holistic and Correct Sentences Scales. Factor analysis assumes that the distribution of the test items corresponds to the structures or variables observed (Kopriva, 2008). Factor analysis is also a procedure that aims to identify patterns in a relatively large set of variables, in order to ensure a tool's construct validity (Hartas, 2010).

For the Holistic Scale, factor analysis showed that the test items were not grouped according to the competences they measured and were classified into 5 components. Some questions, such as Question 1, Part (a) and Question 4's 'past' and 'making a complaint' sections were subsequently omitted because they could have been the reason for an invalid categorization of the questions. The number of extraction was identified by four factors; the questions are not grouped properly in this case. It may be concluded that even if the Holistic Scale achieved inter-rater reliability, it did not prove the validity of the test questions.

For the multiple-trait scale, factor analysis showed that test items were in four categories, but not based on the competences being measured by the test. The items shared by different competences were then omitted, namely Question 1, Part (a), Question 4 – the parts about introducing oneself and making a complaint, and Questions 5 and 6. Moreover, the number of extraction was determined by four factors. The output, is nine items organized appropriately into four components, based on the competences they measure.

It can be concluded that the third hypothesis is accepted, but after omitting some invalid items and determining the number of components as four. After assessing the validity of the test items, their reliability was ascertained to test the fourth hypothesis, namely that '*The test questions assessing communicative competence are reliable*'. Based on the valid questions, classified by the factor analysis of the communicative competence categories, the reliability of the test was computed using Cronbach's alpha. Cronbach's alpha is most suitable, because the test was not repeated, due to the research limitations of access and time, and the difficulty of running parallel tests (Alderson et al., 1995; Fulcher & Davidson, 2007; Jones, 2013). The test's questions are reliable, Cronbach's alpha for wh- questions is (.797), past tense is (.994), and is discourse competence (.803), while for those measuring development in the learners' sociolinguistic competence is (.585). This is based on an interpretation of the reliability index, where if the result is due to unsystematic changes or chance, the reliability index will be close to .00 and a perfect reliability index will be +1.0 (Alderson et al., 1995).

## 5. DISCUSSION AND CONCLUSION

As mentioned in the literature review, there is a shortage in the literature of designed tests and scoring scales to assess FLLs' communicative competence, especially for scoring the learners' competence, rather than their level of proficiency. For this reason, the first question in this study seeks to provide a reliable scoring scale and the second question seeks to present a valid and reliable test.

With respect to the first research question, it was found that the Holistic Scale and Correct Sentences Scale were reliable and the raters were consistent in using them. This result agrees with the findings of Kuiken and Vedder (2014), where the raters were consistent in using a Likert scale- which is a kind of holistic scale. Many reasons underlie this consistency, such as the training received by the raters and the clarity of the scales. Moreover, consistency was measured, because a limited scope of variation in scoring is to be expected, due to the differences between raters (Meier et al., 2006).

Another important finding was derived from a comparison of variations in the means of students' responses for each question and for each scale. The Correct Sentences Scale was more appropriate than the Holistic Scale for scoring grammatical competence past tense questions, where certain components need to be present in the answer. However, this was not required in the sociolinguistic competence questions. On the other hand, the Correct Sentences Scale was not appropriate for scoring sociolinguistic competence questions, while the Holistic Scoring Scale was, because it showed more variation in the learners' responses. This is due to the fact that it rates learners' responses, which mostly contain correct sentences, with few or no incorrect ones and many missing parts. This might also be the reason for low reliability in this competence, as mentioned earlier.

In reviewing the literature, a new scale was designed for this study to fulfil its purpose, similar to East's (2009) new rating scale designs for assessing learners' writing. Moreover, favouring the Holistic Scale in writing assessment is consistent with the findings of Barkaoui (2007), who favours the use of a holistic scale

for this purpose. Finally, the result of the weaknesses and their effects on the test's reliability are in accordance with Rezaei and Lovorn's (2010) confirmation that no research studies have found that the use of a particular scoring scale decreases a test's reliability.

Therefore, this study suggests using the Correct Sentences Scale for scoring questions that measure learners' grammatical competence. Moreover, this study proposes amending this scale so it can also be used for scoring sociolinguistic competence. This is achieved by identifying the necessary components for each response to be scored and dividing the number of components by the score that should be given to the learner for the question (for example a score of 5, as in this study). Through this amendment, this scale becomes a multiple-trait scale which includes the features of a specific task and has different sub-scales for assessing different aspects of the response (Carr, 2011).

With respect to the second research question, it was found that the test items were valid. It was clear that the Correct Sentences Scale was able to categorize the test questions into the competences the test aimed to measure, after omitting the invalid items. Another important finding was that although the items had high validity, especially the sociolinguistic competence questions, they had low reliability. This result may be explained by the fact that the items' instructions and the scoring scale need to be reviewed, which is supported by Rezaei and Lovorn's (2010) finding that achieving high reliability in a writing assessment scale is not easy. Moreover, it is supported by Perkins' argument that holistic scales have low reliability, but high validity (Perkins, 1983).

It can be concluded from this section that these results need to be interpreted with caution, due to the small sample size. The recommended changes should be made to the scale, with the test being administered again. It is also recommended to use the test with a larger number of students and the suggested assessment scale.

## REFERENCE LIST

- Agbatogun, A. O. (2014). Developing learners' second language communicative competence through active learning: clickers or communicative approach? *Educational Technology & Society*, 17(2), 257.
- Alcón Soler, E. (2005). Does instruction work for learning pragmatics in the EFL context? *System*, 33(3), 417-435. doi:10.1016/j.system.2005.06.005
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Ayyanathan, N., Kannammal, A., & Rekha, A. B. (2012). Students' Communicative Competence Prediction and Performance Analysis of Probabilistic Neural Network Model. *International Journal of Computer Science Issues (IJCSI)*, 9(4), 312-317.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.
- Bardovi-Harlig, K., & Vellenga, H. E. (2012). The effect of instruction on conventional expressions in L2 pragmatics. *System*, 40(1), 77. doi:10.1016/j.system.2012.01.004
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, 12(2), 86-107. doi:10.1016/j.asw.2007.07.001
- Bissenbayeva, Z., Ubniyazova, S., Saktaganov, B., Bimagambetova, Z., & Baytucaeva, A. (2013). Communicative Competence Development Model. *Procedia-Social and Behavioral Sciences*, 82, 942-945.
- Brooks, F. B. (1992). Communicative competence and the conversation course: A social interaction perspective. *Linguistics and Education*, 4(2), 219-246. doi:10.1016/0898-5898(92)90015-0
- Brown, J. D. (2001). Pragmatics tests: Different purposes, different tests. *Pragmatics in Language Teaching*, 301326
- Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34(4), 21-38.
- Canale, M. (1987). The measurement of communicative competence. *Annual Review of Applied Linguistics*, 8, 67-84.

- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.
- Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, 6(2), 5-35.
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272. doi:10.1017/S0267190599190135
- Chomsky, N. (1988). *Aspects of the Theory of Syntax*. MIT press.
- Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics*, 22, 263-278.
- East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, 14(2), 88-115. doi:10.1016/j.asw.2009.04.001
- Fischer, R. A. (1984). Testing Written Communicative Competence in French. *The Modern Language Journal*, 68(1), 13-20.
- Fulcher, G. (2013). *Practical Language Testing*. Hoboken: Routledge.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. London: Routledge.
- Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). Problematizing Rating Scales in EFL Academic Writing Assessment: Voices from Iranian Context. *English Language Teaching*, 5(8), 76. doi:10.5539/elt.v5n8p76
- Green, A. (2014). *Exploring language assessment and testing: language in action*. Milton Park, Abingdon, Oxon: Routledge.
- Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9(2), 122-159. doi:10.1016/j.asw.2004.06.001
- Hayen, A., Dennis, R. J., & Finch, C. F. (2007). Determining the intra-and inter-observer reliability of screening tools used in sports injury research. *Journal of Science and Medicine in Sport*, 10(4), 201-210.
- Henning, G. (1987). *A guide to language testing: development, evaluation, research*. New York; London: Newbury House.
- Hirdes, J. P., Prendergast, P., Smith, T. F., Morris, J. N., Rabinowitz, T., Ikegami, N., . . . Curtin Telegdi, N. (2002). The Resident Assessment Instrument\_Mental Health (RAI\_MH): Inter-Rater Reliability and Convergent Validity. *The Journal of Behavioral Health Services & Research*, 29(4), 419-432. doi:10.1097/00075484-200211000-00006
- Hymes, D. (1962). The ethnography of speaking. *Anthropology and Human Behavior*, 13(53), 11-74.
- Hymes, D. (1972). On communicative competence. *Sociolinguistics*, 269-293.
- Jones, N. (2013). Reliability and dependability. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 350-362) Routledge.
- Kenyon, D., & McGregor, D. (2013). Pre-operational testing. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 295-306) Routledge.
- Knoch, U. (2009a). *Diagnostic writing assessment: the development and validation of a rating scale*. Frankfurt am Main: Peter Lang.
- Knoch, U. (2009b). *Language Testing and Evaluation, Volume 17: Diagnostic Writing Assessment : The Development and Validation of a Rating Scale*. Peter Lang AG.
- Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2), 81-96. doi:10.1016/j.asw.2011.02.003
- Kopriva, R. J. (2008). *Improving testing for English language learners*. New York: Routledge.

- Lukin, A. (2013). Journalism, ideology and linguistics: The paradox of Chomsky's linguistic legacy and his 'propaganda model'. *Journalism*, 14(1), 96-110.
- McNamara, T. (1996). *Measuring second language performance: a new era in language testing*. London: Longman.
- Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: factors related to discrepancies in scoring. *Assessment in Education*, 13(1), 1.
- Perkins, K. (1983). On the Use of Composition Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability. *TESOL Quarterly*, 17(4), 651-671.
- Pillar, G. W. (2012). *A Framework for Testing Communicative Competence*. (). Hungary: University College of Nyiregyhaze:
- Plakans, L. (2013). Writing Scale Development and Use Within a Language Program. *TESOL Journal*, 4(1), 151-163. doi:10.1002/tesj.66
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39. doi:10.1016/j.asw.2010.01.003
- Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, 16(4), 457-478. doi:10.1177/026553229901600403
- Schachter, J. (1996). Chapter 5 - Maturation and the Issue of Universal Grammar in Second Language Acquisition. *Handbook of Second Language Acquisition* (pp. 159-193). San Diego: Academic Press. doi:<http://dx.doi.org.ezp.ud.edu.sa/10.1016/B978-012589042-7/50007-4>
- Strakšienė, G., & Baziukaitė, D. (2010). Development of Children's Communicative Competence using Case Method of Technology. (pp. 253-258). Dordrecht: Springer Netherlands. doi:10.1007/978-90-481-3656-8\_47
- van Loo, M. A., Moseley, A. M., Bosman, J. M., de bie, R. A., & Hassett, L. (2003). Inter-rater reliability and concurrent validity of walking speed measurement after traumatic brain injury. *Clinical Rehabilitation*, 17(7), 775-779. doi:10.1191/0269215503cr677oa
- Van Schaik, E., Lynch, E. M., Stoner, S. A., & Sikorski, L. D. (2014). Can a web-based course improve communicative competence of foreign-born nurses? *Language, Learning & Technology*, 18(1), 11.
- Wang, P. (2009). The Inter-Rater Reliability in Scoring Composition. *English Language Teaching*, 2(3), 39. doi:10.5539/elt.v2n3p39
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Weir, C. J. (1990). *Communicative language testing*. New York; London: Prentice Hall.
- Williams, E. (1979). Elements of Communicative Competence. *English Language Teaching Journal*, 34(1), 18.
- Yano, Y. (2002). Communicative competence and English as an international language. *Intercultural Communication Studies*, 12(3), 75-83.
- Young, R. (2013). Social dimensions of language testing. In G. Fulcher, & F. Davidson (Eds.), *The routledge handbook of language testing* (pp. 178-193) Routledge.

## APPENDICES

### Appendix 1: The Scoring Scales

Table 1. The Holistic Scoring Scale

The mark	The score descriptors
4	The student's response is appropriate and relevant to the task set. It shows that she is competent in the task.
3	The student's response is mostly related to the task, although there are a few gaps in the

	answer. It shows that she is nearly competent in the task.
2	The student's response is mostly related to the task, though there are many gaps in the answer. It shows that she is somewhat competent in the task.
1	The student's response is limited in relation to the task set. It shows that she has almost no competence in the task.
0	The student's response is not appropriate to the task set. It shows that she is not competent in the task.

Table 2. The Correct Sentences Scale

The question	The criteria	The mark (out of 5)	The score descriptors
1	<b>Scale 1: To assess 'asking wh-questions'</b> 1 mark for each of the question components	5 marks	5 shows that she is competent in the task 4 shows that she is nearly competent in the task 3 shows that she is somewhat competent in the task
2, 3, 4, 5 and 6	<b>Scale 1: To assess discourse competence</b> 1 mark for using a pronoun correctly 1 mark for using a conjunction correctly	1 mark *5 1 mark *5	2 shows that she has almost no competence in the task 1 shows weaknesses in her competence in the task
	<b>Scale 2: To assess 'writing a passage in the past' and sociolinguistic competence:</b> Count the number of correct and appropriate sentences Count the number of incorrect and inappropriate sentences	Divide the number of correct sentences by the total of correct and incorrect sentences. The result is multiplied by 5.	0 shows that she is not competent in the task