Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 12, No. 2, October 2022
ISSN 2476-5880

Psychometric Evaluation of Dictations with the Rasch Model

Rasha Abed Hussein[1], Shaker Holh Sabit[2], Merriam Ghadhanfar Alwan[3], Hussam Mohammed Wafqan[4*], Abeer Ameen Baqer[5], Muneam Hussein Ali[6], Safa K. Hachim[7,8], Zahraa Tariq Sahi[9], Huda Takleef AlSalami[10], Bahaa Aldin Fawzi Sulaiman[11]

Received: April 2022 Accepted: May 2022

Abstract
Dictation is a traditional technique for both teaching and testing overall language ability and listening comprehension. In a dictation, a passage is read aloud by the teacher and examinees write down what they hear. Due to the peculiar form of dictations, psychometric analysis of dictations is challenging. In a dictation, there is no clear boundary between the items and every word in the text is potentially an item. This makes the analysis of dictations with classical and modern test theories rather difficult. In this study, we suggest a procedure to make dictations analyzable with psychometric models. Our strategy entailed using several independent short passages instead of a single long passage. The number of mistakes in each passage was counted and entered into the analysis. Rasch model analysis was then applied to the passage scores (mistakes). Our findings showed that dictations fit the Rasch model very well and it is possible to measure examinees' ability on an interval scale using dictations.

Keywords: dictation, partial credit model, Rasch model, reduced redundancy tests, validation

1. Introduction

Dictation is a very old teaching and testing technique, especially in French classes, and is referred to as dictee in French (Valette, 1964). Dictation is a kind of reduced redundancy test (RRT; Spolsky et al., 1968). In an RRT, parts of a linguistic message are covered and examinees

[1]Al-Manara College for Medical Sciences (Maysan)/Iraq
[2]Scientific Research Center, Al-Ayen University, Thi-Qar, Iraq
[3]Medical Laboratory Techniques Department /Medical (Technology) College, Al-Farahidi University/Iraq
[4]English Language Department, Al-Mustaqbal University College, Babylon, Iraq hussam.mohammed@mustaqbal-college.edu.iq
[5]Medical Laboratory Techniques Department, Dijlah University College, Baghdad, Iraq
[6]Al-Nisour University College/ Iraq
[7]College of technical Engineering, The Islamic University, Najaf, Iraq.
[8]Medical Laboratory Techniques Department, Al-Turath University College, Iraq
[9]Department of Dentistry, Al-Zahrawi University College, Karbala, Iraq
[10]Altoosi University College, Najaf , Iraq
[11]Al-Esraa University College, Baghdad, Iraq

are required to process the incomplete message. The rationale is that proficient users of the language can understand the message when parts of it are deleted. Examples of RRT are cloze test, C-Test, noise test, etc. Reduced redundancy in dictation is operationalized by using a written text and presenting it orally. Using this strategy, the examinee is deprived of referring to the passage when needed (Klein-Braley, 1994).

Dictation procedure was refuted by psychometric-structuralists scholars as an ineffective testing technique that does not measure any important language skills or components because everything is given to examinees (Klein-Braley, 1997; Oller, 1979). According to Lado (1961), dictation does not measure vocabulary and syntax and it was even considered a poor test to assess listening comprehension. Oller (1971) criticized Lado's view and pointed out the complexity of the dictation and argued that in dictation an active process of analysis is required. Oller observed high correlations between dictation and other language tests such as vocabulary, grammar, composition, and phonology and found it as a valid measure of English language skills.

However, for a long time applied linguists did not have a positive attitude toward dictation. For example, Lado (1961) stated that dictation does not measure language because in a dictation everything is given, i.e., the words and the word order. It does not test aural perception because they can understand words from context and students can hear the words when the text is read slowly which is common in dictation. Harris (1969) says that as a testing technique it is uneconomical and inaccurate. Anderson (1953) states that it is an inadequate and indirect test of listening comprehension which is a very important skill. Somaratne (1957) states that it is only a test of spelling.

Nevertheless, Oller (1971) disagrees with the above statements and argues that the word order is only available for the speaker or the examiner who reads the text aloud. Oller refers to Saussure (1959) who wrote that the sound chain we hear is linear; it is a continuous line without any apparent or noticeable division. Oller (1971) argued that in order to divide the chain of speech, the listener or the examinee in this case should get involved in analysis by synthesis. His analysis of the errors made by language learners while taking dictations showed that they had made several word order mistakes. This was used as evidence that word order is not given in dictation and examinees have to extract it from the speech chain. The other errors were incorrect words or phrases and writing words that were not in the text. Oller (1971) stated that examinees have to receive the auditory input and process it to reproduce the original sentence that is read to them. This is not the simple task that Lado and others claim dictation is.

2. Review of Literature

Empirical studies on dictation also support Oller's arguments about dictation and refute the statement of others who state that dictation is a simple task. In an experimental research study, Valette (1964) showed that dictation improved French students' sentence construction ability but has no effect on their other language skills. She also showed that dictee correlates highly with learners' final French exam scores in two different groups (r=.78 & r=.89) and suggested that dictee can be used as a valid test of overall language ability. However, she warns that if dictee is

used regularly in the classroom context, then examinees' scores are the result of practice and do not necessarily show their proficiency level. In a study on students of German as a foreign language, Valette (1967) found a correlation of .90 between dictation and combined scores of writing, listening, and reading. Oller (1971) also showed very high correlations between dictation and the vocabulary, grammar, writing, and pronunciation sections of a placement test at the University of California. Oller (1971) showed that dictation correlates at .86 with the combined score of the criterion measures. Irvine, Atai, and Oller (1974) showed that dictation correlated with the TOEFL overall score at r=.69 and with its subsections: listening (r=.69), structure (r=.63), vocabulary (r=.47), reading comprehension (r=.53), and writing (r=.52). Dictation also had a high correlation with the cloze test administered along with the TOEFL (r=.75).

More recently, Yazdinejad and Zeraatpishe (2019) used a variation of dictation proposed by Johansson (1973) called partial dictation in English as a foreign language context in Iran. In a partial dictation, examinees listen to a passage that is tape recorded while they have the transcribed written text of the recording with some presumably difficult parts deleted. Examinees have to listen to the recording and fill in the missing parts in the written text. Their study showed that partial dictation is reliable and correlates highly with a cloze test, a C-Test, and a reading comprehension test. Exploratory factor analysis showed that partial dictation loads on a general language proficiency factor with other tests.

In this study, we examine the psychometric properties of dictations with the Rasch model. Due to the peculiar form of dictations, their analysis with IRT models is challenging. The challenge is that in dictations the line between individual items is not clear and one does not know what should be entered into the IRT analysis as individual items. In this study, we suggest a procedure to make dictations analyzable with IRT models. We specifically analyze a set of dictations with the Rasch partial credit model (PCM; Masters, 1982) and examine the fit of the data to the PCM. Following the C-test literature, we suggest using several short passages instead of a single passage for dictation. Then each passage can be entered into the analysis as an item or the unit of analysis. Recently, Dhyaaldian, Al-Zubaidi, et al. (in press) used the same procedure for the analysis of cloze tests with the Rasch model. However, in dictations, the number of errors or mistakes in each passage should be entered as passage scores. This is similar to Rasch's study (1960/1980) in the analysis of oral readings where examinees read passages and the number of errors was counted and used for analysis. Since in dictations errors are modeled instead of correct answers, higher scores (i.e. errors) for persons mean lower ability and higher scores (errors) for items mean harder items. This is the opposite of the standard procedure of modeling correct answers where higher scores for persons mean higher ability and higher scores for items mean easier items. To solve this problem, the orientation of the scale was reversed in WINSTPES by setting USCALE=-1. This is equivalent to simply reversing the signs of the item and person parameters.

3. Method

3.1 Participants

A sample of 182 undergraduate Iraqi students (103 female) studying English at Al-Nisour University College in Baghdad took the dictations. Their age ranged from 21 to 36 (M=25.85, SD=4.21).

3.2 Instrument

To construct the dictation, reading compression passages from the British Council website were used (https://learnenglish.britishcouncil.org/). The reading comprehension exercises on the British Council website are presented in five levels A1, A2, B1, B2, and C1. For the purpose of this study, three passages from the B1 level were selected and three passages from the B2 level were selected. Since the participants in this study were lower intermediate and intermediate learners of English, texts from other levels were not considered. To ensure uniformity and for the sake of standardization, the passages were read aloud by a near-native English teacher and were recorded on CD. The length of the passages ranged from 92 to 156 words.

3.3 Procedures

The recording was played for the students as a mid-term exam in a listening comprehension course in six parallel classes. They were instructed to write down whatever they hear. The written texts produced by the examinees were corrected by the researchers and the number of errors in each passage was counted. To provide validity evidence for the dictation, the Rasch measurement model (Rasch 1960/1980) was used. As explained before, to solve the problem of lack of clear boundaries between the items in dictation, the number of errors in each passage was entered into the analysis. Masters' (1982) partial credit model which accommodates polytomous items with a different number of response categories was used to analyze the data. The PCM is the appropriate model for analyzing dictations because the rating scale model (RSM; Andrich, 1978) requires equal number of response categories in each item. In dictation, with a different number of words in each passage, the number of response categories is not defined. The number of response categories could be equal to the number of words in each passage and each word is a potential item. Thus, PCM which does not require an equal number of response categories or equal distances between the thresholds across items is the appropriate model for analyzing dictations. Winsteps Rasch model computer program version 5.2.2 (Linacre, 2022a) was used to perform the analyses.

4. Results and Discussion

Table 1 shows the item difficulty parameters, their standard errors, and their infit and outfit mean square values. Infit and outfit mean square statistics show the amount of randomness or distortion in the response patterns. High values of infit and outfit statistics indicate that the data do not follow the pattern expected by the Rasch model (Linacre, 2002). The expected value is 1 but infit and outfit mean square values in the range of .60 to 1.40 for polytomous items are acceptable

(Bond & Fox, 2007). Infit mean square values show that Item 4 with an infit mean square value of 1.42 misfits the model and should be discarded. The other items fit the model perfectly. It should be noted that the appealing properties of the Rasch model including unidimensionality and interval scaling are only achieved if the data fit the model (Baghaei et al., 2017). Point-measure correlation is equivalent to item-total correlation in the classical test theory. However, here instead of correlating items with the total raw scores, they are correlated with the person ability parameters. Higher point-measure correlations indicate a stronger relationship between the item and performance on the overall test (item discrimination).

Table 1
Item measures and fit statistics for the six dictation passages

Item	Diff.	SE	Infit MNSQ	Outfit MNSQ	Pt. Meas. Cor.
1	-.06	.04	.92	.97	.89
2	.01	.04	.99	.92	.89
3	-.13	.04	.88	.90	.89
4	-.03	.04	1.42	1.25	.86
5	-.03	.04	.96	.92	.90
6	.24	.04	.78	.86	.90

Note. Diff=Difficulty Parameter; SE=Standard Error; Pt. Meas. Cor. =Point-Measure Correlation

Table 2 shows sample and test statistics. According to Table 2, the reliability of the dictation test battery with six items (considering each passage as a polytomous item) is very high (r=.91) which shows that the examinees have been measured very precisely. A person separation value of 3.28 shows that the test can identify at least three levels of ability strata. Item separation value of 2.67 shows that respondents have identified more than two levels of difficulty strata in the items. The lower value for item separation means that the dictation passages were close in difficulty which is also evident from their difficulty measures in Table 1. Principal components analysis (PCA) of standardized residuals showed that the eigenvalue in the first contrast after applying PCA to residuals is 1.5. This is an indication that the strength of the first component derived from the residuals is very small and, therefore, the data are unidimensional (Baghaei & Cassady, 2014; Linacre, 2022b).

Table 2
Overall test and sample statistics

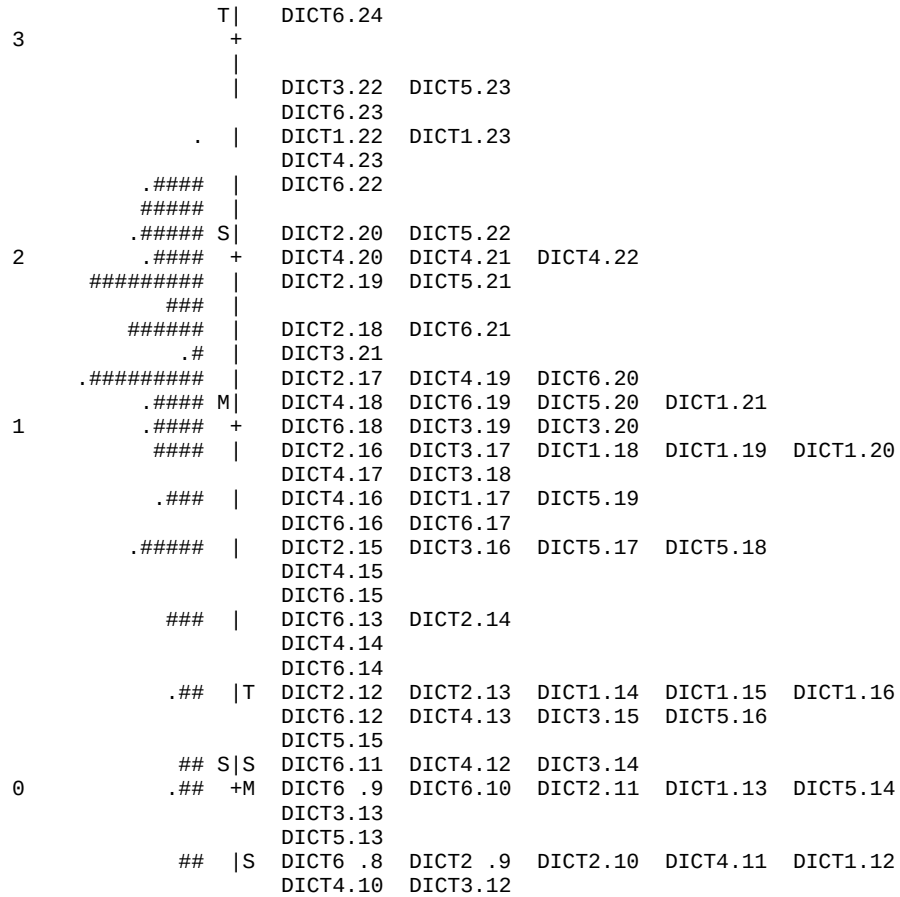
Reliability	.91
Person separation	3.28
Item separation	2.67
Mean (SD)	1.12 (.98)
Range	5.21
Eigenvalue in 1st Contrast	1.50

 ISSN 2476-5880

Figure 1 is the Wright map for the items and persons. The map shows that the items and their thresholds cover a wide range of the scale. The majority of persons and items are in front of each other which means that the test is well-targeted for the sample. The item thresholds cover the entire span of person abilities. That is, there are no examinees outside the range where the item thresholds fall. This is evidence that the test can precisely measure all examinees at all levels of ability.

5. Conclusion

Dictation is a traditional testing and teaching technique that originated in the French classes (Valette, 1967). It is a member of the family of reduced redundancy tests in which parts of the message are masked and examinees are required to process the language (Klein-Braley, 1997; Rasoli, 2021). At the beginning of the psychometric-structuralist era of language testing, dictation was rejected as a futile testing technique which at best measures only spelling (Lado, 1961). However, with the rise of the integrative approach to testing, dictation was given new attention and was advocated as an overall test of general language proficiency (Oller, 1979). Recently, a variation of dictation called partial dictation (Johansson, 1973) has been successfully used for testing listening comprehension in English as a foreign language (Cai, 2012).



 ISSN 2476-5880

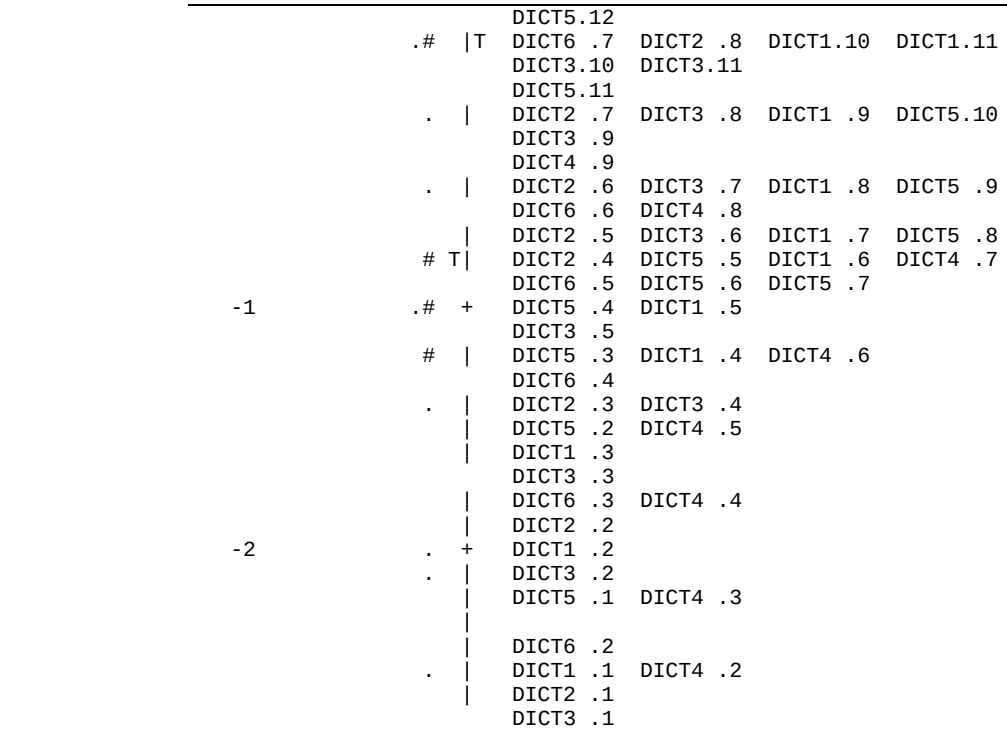


Figure 1
Item-person map

Analysis of dictations within the classical and modern test theory frameworks is challenging. This is due to the fact that in dictations, it is very hard to identify individual items. In dictation, examinees listen to a text that is read aloud to them and try to reproduce the text in its entirety by writing the words in the exact form and order they hear. The final product is a text which is not suitable to be analyzed with psychometric technique because it does not have distinct items.

In this study, we suggested a procedure to make dictation analyzable with the Rasch measurement model. We followed the procedure used for the analysis of C-Tests (see Forthmann et al., 2020) by using several short independent dictation passages instead of a single long passage. Each dictation was then treated as a polytomous item by entering the number of errors on each passage into the analysis as the item raw score. Masters' (1982) PCM was then used to scale the test and the passages. The findings showed that the analysis of dictations with the Rasch model is possible. All the passages except one fitted the Rasch model. The test was highly reliable and unidimensional. Examination of the Wright map showed that the texts cover a wide range of the ability and target the persons' abilities accurately. These findings show that dictations are scalable with the Rasch model and constructing a unidimensional interval scale of ability is possible when examinees are tested with dictations.

In this research, we entered the number of errors in each passage as the item score. Future research could examine other scoring strategies such as the total number of words in a passage

minus the number of errors and compare the fit and other psychometric test properties with those obtained when the number of errors alone are used as passage scores (for a complete review of scoring strategies that could be applied to tests like dictation see Baghaei et al., 2019 and Nadri et al., 2019). If dictation is administered in a speeded mode, by reading the passage only once at the normal pace, the Rasch Poisson Counts Model (RPCM; Rasch, 1960/1980) can be used for modeling the data. Baghaei and Doebler (2019) provided a complete review of RPCM with R codes and applied it to a speeded test. Polytomous multidimensional Rasch models (Adams et al., 1997) can be used for modeling dictations composed of different types of passages that are deemed to be multidimensional (Bagahei, 2013; Baghaei & Grotjahn, 2014). Other polytomous Rasch and IRT models such as the graded response model (Samejima, 1969) and continuous response model (Samejima, 1973) can be used for modeling dictations and the findings may be compared with those of the present study (see Dhyaaldian, Kadhim, et al., in press). In this research, we only examined intermediate texts for intermediate university English language learners. Future research should evaluate the suitability of dictation for examinees at other levels of proficiency and in other languages.

References

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Anderson, D. F. (1953). Tests of achievement in the English language. *English Language Teaching, 7*, 37-69.

Bond, T. G., & Fox, C. M. (2007,). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd Ed.). New York: Lawrence Erlbaum.

Baghaei, P., & Doebler, P. (2019). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports*, *122* (5), 1967-1994.

Baghaei, P., Ravand, H., & Nadri, M. (2019). Is the d2 test of attention Rasch scalable? Analysis with the Rasch Poisson Counts Model. *Perceptual and Motor Skills*, *126*, 70-86.

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, *19*, 155-168.

Baghaei, P., & Cassady, J. (2014). Validation of the Persian translation of the Cognitive Test Anxiety Scale. *Sage Open, 4*, 1-11.

Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Rasch Model. *Psychological Test and Assessment Modeling, 56*, 60-82.

Baghaei, P. (2013). Development and psychometric evaluation of a multidimensional scale of willingness to communicate in a foreign language. *European Journal of Psychology of Education, 28*, 1087-1103.

Cai, H. (2012). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing, 30*, 177-199.

Dhyaaldian, S. M. A., Al-Zubaidi, S. H., Mutlak, D. A., Neamah, N. R., Albeer, M. A., Hamad, D. A., Al Hasani, S. F., Jaber, M. M., & Maabreh, H. G. (in press). Psychometric evaluation of cloze tests with the Rasch model. *International Journal of Language Testing, 12* (2).

Dhyaaldian, S. M. A., Kadhim, Q. K., Mutlak, D. A., Neamah, N. R., Kareem, Z. H., Hamad, D. A., Tuama, J. H., Qasim, M. S. (in press). A comparison of polytomous Rasch models for the analysis of C-Tests. *International Journal of Language Testing, 12* (2).

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-tests. *Journal of Psychoeducational Assessment*, *38*(6), 692-705.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Irvine, P., Atai, P., & Oller, J. W. Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24,* 245–252.

Johansson, S. (1973). *The partial dictation as a test of foreign language proficiency.* Swedish-English Contrastive Studies, Report No. 3.

Linacre, J. M. (2022a). *Winsteps® Rasch measurement computer program (Version 5.2.2)*. Portland, Oregon: Winsteps.com.

Linacre, J. M. (2022b). *Winsteps® Rasch measurement computer program User's Guide*. *(Version 5.2.2)*. Portland, Oregon: Winsteps.com.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*, 47-84.

Klein-Braley, C. (1994). *Language testing with the C-Test. A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty.* Unpublished post-doctoral thesis (Habilitations-schrift), Universität-Gesamthochschule Duisburg: Fachbereich 3, Sprach- und Literaturwissenschaften.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. Bristol, Inglaterra: Longmans, Green and Company.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16* (2), 878.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Nadri, M., Baghaei, P., & Zohoorian, Z. (2019). Analysis of the Ruff 2 & 7 Test of Attention with the Rasch Poisson Counts Model. *The Open Psychology Journal*, *12*, 7-11.

Oller, J. W. Jr. (1971). Dictation as a device for testing foreign language proficiency. *English Language Teaching, 15,* 254–259.

Oller, J. W. Jr. (1979). *Language tests at school*. London: Longman.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (expanded Ed.). Chicago, IL: University of Chicago Press.

Rasoli, M. K. (2021). Validation of C-test among Afghan students of English as a foreign language. *International Journal of Language Testing, 11*(2), 109-121.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203-219.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika,* Monograph Supplement No. 17.

Saussure, F. D. (1959). *Course in general linguistics*. New York: Philosophical Library.

Somaratne, W. (1957). *Aids and tests in the teaching of English*. Oxford: Oxford University Press.

Spolsky, B., Sigurd, B., Sato, M., Walker, E., & Arterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning, 18* (Suppl 3), 79–101.

Valette, R.M. (1964). The use of the dictée in the French language classroom. *The Modern Language Journal*, *48*, 431-434.

Valette, R.M. (1967). *Modern language testing*. New York: Harcourt Brace and World.

Yazdinejad, A., & Zeraatpishe, M. (2019). Investigating the validity of partial dictation as a test of overall language proficiency. *International Journal of Language Testing, 9,* 44-56.