Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Vol. 9, No. 2, October 2019

Investigating the Validity of Partial Dictation as a Test of Overall Language Proficiency

Anoushe Yazdinejad[1], Mitra Zeraatpishe[2]

Received: 12 December 2018 Accepted: 17 February 2019

Abstract

In this study the validity of partial dictation as a measure of overall language proficiency was examined. Two partial dictation tests along with a C-Test, a cloze test, and a reading comprehension test, as criterion measures, were administered to a group of Iranian EFL learners. The coefficients of correlation between partial dictation and criterion measures were computed. Correlations revealed that partial dictation highly correlates with the cloze test, the C-Test, and the reading comprehension test. Principal components analysis showed that all the variables (four C-Test passages, two dictation passages, one cloze passage, and one reading comprehension passage) formed one single factor which explained 58% of the variance. All the variables had high loadings on the factor. These findings were interpreted as evidence of the similarity of the construct measured by partial dictation and other measures employed in the study. Desiderata for future research and potential applications of partial dictation in foreign language testing and teaching are discussed.

Keywords: Dictation, partial dictation, C-Test, reduced redundancy tests, validation

1. Introduction

Reduced redundancy principle (RRP) is a theory proposed by Spolsky (1968) to account for the noise test that he had developed to measure general language proficiency. According to the reduced redundancy principle languages contain a lot of extra information that is redundant. Redundancy is present in our formal and informal spoken and written language. Normal speech and writing contain extra information for emphasis or clarity. The reduced redundancy principle states that a proficient user of a language should be able to comprehend the language when redundancy is eliminated. A piece of writing with some missing words should be comprehensible to proficient users of the language. Furthermore, one should be able to understand a speaker when s/he is talking in a busy street with lots of noise. One's knowledge of language and the redundancy in the language help proficient users to overcome noise that is introduced into the communication system.

RRP has been employed to account for a number of tests including cloze test, noise test, C-Test, cloze elide, dictation, and partial dictation among others. The noise test is a kind of dictation in which hissing noise is imposed on orally presented sentences and examinees

[1] *English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran.*
[2] *English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran. Corresponding author: Email: mitra.zeraatpishe@yahoo.com*

are expected to listen and write the sentences (Gaies, Gradman, & Spolsky, 1977). In cloze test, every n^{th} word in a relatively long passage is deleted (*n* is between 5 to 9) and examinees are expected to fill in the missing words (for variations of the cloze test see Oller, 1979). Cloze test was later criticized for indeterminacy of the construct (Klein-Braley & Raatz, 1984) and recently for its method factor that contaminates the test scores (Baghaei & Ravand, 2016; Baghaei & Ravand, 2019; Sheybani & Zeraatpishe, 2018).

The C-Test was proposed as a replacement for the cloze test to overcome its shortcomings. In C-Test, the second half of every second word is deleted in a number of independent short passages. That is, in a standard C-Test the rate of deletion is two but half of the word is deleted (Klein-Braley & Raatz, 1984). C-Test is a very well-researched RRP test and over the past decades different variations of the C-Test have been proposed (Baghaei & Grotjahn, 2014a; Baghaei & Grotjahn, 2014b; Baghaei, 2014; Baghaei, 2010a; Eckes & Grotjahn, 2006; Fadaeipour & Zohoorian, 2017; Grotjahn, Schlak, & Aguado 2010).

In a cloze elide test, irrelevant words are randomly inserted in a passage and examinees are required to identify them. The ability of the examinees to recognize the redundant words is considered to be related to their level of language proficiency. That is, the more proficient language learners, the more successful they will be in identifying the irrelevant words in the text (Holster, 2017; Manning, 1987; Zare & Boori, 2018). Dictation and partial dictation are explained in more detail below. In all these test types parts of the written or spoken message is masked and examinees are expected to process the language. The more successful the examinees are in comprehending the language under the reduced redundancy condition the more proficient the examinee is deemed to be (see Baghaei, 2011 for an overview).

2. Review of Literature

2.1 Dictation

As a member of the family of RRP tests, dictation has a long tradition in the French language classes and was routinely used as a technique for both learning and testing (Valette, 1964). For many years scholars argued that dictation is not a valuable technique for testing foreign languages. For example, Lado (1961) stated that:

> "Dictation…on critical inspection…appears to measure very little of language. Since the word order is given…it does not test word order. Since the words are given…, it does not test vocabulary. It hardly tests the aural perception of the examiner's pronunciation, because the words can in many cases can be identified by context…the student is less likely to hear the sounds incorrectly in the slow reading of the words which is necessary for dictation" (p. 34).

Along the same line, Harris (1969) stated that "As a testing device…dictation must be regarded as generally both uneconomical and imprecise" (p. 5). Anderson (1953) wrote that "Some teachers argue that dictation is a test of auditory comprehension, but surely this is a very indirect and inadequate test of such an important skill" (p. 43). Somaratne (1957) argued that "Dictation is primarily a test of spelling" (p. 48).

However, Oller (1971) refutes these claims and mentions that Lado's statement that the word-order is given in dictation is only correct from the perspective of the speaker/examiner who knows the word order. He resorts to Saussure's (1959) argument on

speech who wrote "…the main characteristic of the sound chain is that it is linear. Considered by itself it is only a line, a continuous ribbon along which the ear perceives no self-sufficient and clear-cut division… (p. 103-4).

Oller (1971) states that in order to segment the chain of speech an active process of analysis-by-synthesis is required. He gives several examples of dictations errors committed by language learners to support his argument that in dictation the word order is not given to the test takers and they must extract the intended word order from the sequence of sounds they hear. His error examples include order-inversion, incorrect understanding of words or phrases, and insertion of extra words. "The student not only receives auditory information, but he processes this information in order to generate a sentence (or sequence of them) that has meaning. This is by no means the simple activity that Lado's statement implies" (p. 257).

Empirical evidence accumulated over the past five decades supports Oller's ideas about dictation. In an experiment on the role of dictée (dictation) in teaching French, Valette (1964) demonstrated that dictée helped students to score higher on written sentences section of their final exam. Nevertheless, it had no effect on French learners' other language abilities. She also demonstrated that dictée correlated highly (r=.78 and r=.89 in two groups) with learners' final examinations and recommended dictée as a valid overall test of French knowledge provided that students do not have regular practice in dictée because then the scores are affected by practice. In another study, Valette (1967) reported a correlation of .90 between dictation and combined scores of listening, reading, and writing in German.

Oller (1971) demonstrated that dictation correlated highly with the total score and individual components (vocabulary, grammar, composition, phonology) of the English as a Second Language Placement Examination for the University of California at Los Angeles. The correlation between the total score and the dictation was .86. Irvine, Atai, and Oller (1974) showed that dictation correlated with TOEFL (Test of English as a Foreign Language) total scores (r=.69) and its components: listening comprehension (r=.69), English structure (r=.63), vocabulary (r=.47), reading comprehension (r=.53), and writing ability (r=.52). It also correlated highly with cloze test scored with acceptable word procedure (r=.75).

2.2 Partial Dictation

Johansson (1973) argued that dictation is only appropriate for beginning students and should be modified for more advanced learners. He proposed partial dictation to remedy this problem. In partial dictation, test takers are provided with a passage in which certain presumably difficult portions are blanked out. The passage is tape recorded in its entirety and presented to the test takers acoustically only once. They are required to listen to the recording and write down the missing portions on the written passage.

Johansson (1973) states that partial dictation has some advantages compared to ordinary dictation: (1) It is tape recorded and, therefore, a variety of voices, speech situations, registers, and dialects can be used, (2) it is more economical since it can be administered more quickly compared with ordinary dictation and one can focus on the problematic portions of the text, and (3) the testing situation in partial dictation is more natural since the speech is not interrupted as in the ordinary dictation. To increase the difficulty of partial dictation, Johansson (1973) suggests to select recordings in unfamiliar dialects or with difficult words and constructions or increase the speed of delivery. He further argues that "…aural perception is not a passive skill but involves the active use of all aspects of the language system" (p. 8).

In an empirical study on partial dictation, Johansson (1973) administered two English passages to a sample of Swedish students (*n*=80). The correlation between the scores on the two passages was .80 which was considered as evidence of parallel-form reliability. He reported high correlations between partial dictation and other language measures: English vocabulary (r=.80), English vocabulary + English grammar (r=.86), English pronunciation (r=.78), Swedish vocabulary (r=.65), and English speaking (r=.62). These correlations were used as validity evidence for partial dictation as a measure of language proficiency. Johansson's (1973) further analysis of errors committed by learners who took the partial dictation revealed that "…the student must use his knowledge on all levels of the language system in order to produce a correct response" (p. 40).

Cai (2012) examined the construct validity of partial dictation as a measure of listening comprehension in English as a foreign language. The aim of the study was to demonstrate that partial dictation items can measure higher-order abilities in listening comprehension. He devised two types of dictation items. The first type was a standard partial dictation test. Learners listened to a recording while they had the written form with gaps. They had to fill in the gaps with one word. In the second type, longer chunks were missing and examinees had to supply four-word phrases into the gaps. He also administered two types of constructed response items. In the first type, examinees listened to a recording after which they read 10 statements paraphrasing the key points of the recording. Each statement had a blank to be filled in by one word. In the fourth section, examinees listened to a recording after which they read 10 incomplete statements paraphrasing the key points of the recording which had to be filled in by responses ranging in length from one word to a short clause. The partial dictation items were considered as tests of lower order abilities and the constructed response items as tests of higher order abilities. Confirmatory factor analysis showed that a bifactor model in which all the items formed a general factor and the four item types formed four method specific factors had the best fit compared to a simplex model and a higher order model. This was interpreted as the capability of partial dictation to tap into examinees' higher order abilities as the partial dictation items combined well with the constructed response items to from a general listening factor in the bifactor model.

Sigott (2004) argued that whether partial dictation is a true reduced redundancy test is doubtful since deletions are not on a random basis and test developers decide which portions are 'hard' and should be deleted. He further states that since deletion in partial dictation is not done randomly, Johansson's (1973) claim that partial dictation is a test of general language proficiency should be put to more rigorous empirical tests. In this study, we aim to provide criterion-related validity evidence for partial dictation. The following research questions were formulated:

1. Is there any correlation between partial dictation and reading comprehension?
2. Is there any correlation between partial dictation and multiple-choice cloze test?
3. Is there any correlation between partial dictation and C-test?
4. To what extent partial dictation is valid and reliable?

3. **Methodology**

3.1 Participants and Setting

In order to collect the required data, 112 Iranian undergraduate English as a Foreign Language students participated in the study. They were of different ages, gender and various levels of language proficiency who studied in the Islamic Azad University of Mashhad in the

field of Teaching English. The participants were all non-native speakers whose first language was Persian. The participants included 83 female and 29 male students. They ranged between 19 and 60 years in age (M= 23.86, SD= 7.61).

3.2 *Instrumentation*

In order to collect data for the purposes of the present study, four instruments were employed. The instruments consist of a partial dictation test, a C-test, a cloze test, and a reading comprehension test. The reliability of the instruments were examined.

3.2.1 Partial dictation

To develop the parial dicatation items, two standard written cloze test passages from the past papers of the FCE (2016) (First Certifciate in English) developed by the Cambridge Examinations were employed. The two cloze passages each contained eight blanks. The complete passages were read aloud by a nonnative speaker teacher with native-like proficieny and his voice was recorded on CD. The recordings were played for the participants while they had the gapped passages. Their task was to fill in the gaps by colsely listeing to the recordings and focusing on the missing parts of the texts. The Cronbach's alpha reliability of the partial dicataion with 18 items was .81.

3.2.2 Cloze Test

One multiple-choice cloze test was employed from the Cambridge First Certificate in English, (FCE) past papers. The test contained eight gaps. The passage was about geneology. For each gap there were four words as alternatives and participants were required to choose the correct one as an answer to each gap. Five minutes were allocated for the cloze test. The Cronbach's alpha reliability of the cloze test was .59.

3.2.3 C-Test

The C-Test contained four passages which were developed by Babaei and Shahri (2010). Each passage contained 20 gaps. It was developed by removing the second half of every second word in a text (Babaei & Shahri, 2010). The participants were asked to restore the missing words in the texts by completing the mutilated words. The texts were represented under four topics: Fly in the Airplanes, How to Lose Weight and One Man in a Boat, and The Best Art Critics. The time which was given to each text was 5 minutes. The Cronbach alpha internal consistency of the C-Test was .86.

3.2.4 Reading comprhension

A multiple-choice reading comprehension test from the TOEFL pratcice materials was utilized for the current study. It contained only one passage and 12 multiple choice items from the *TOEFL in Flash* by Broukal (2002). The topic of the passage was about a poet named Horace Pippin. It consisted of 12 multiple choice items. The participants had to read the text and answer the questions in 10 minutes. The Cronbach's alpha reliability of the reading comprehension test was .60.

4. Analysis and Results

Table 1 shows the means, standard deviations, variances, minimum, and maximum for each of the variables in the study. Since the nature and the number of items in each test are different we cannot directly compare the tests. For scoring the tests the number of correct replies were counted.

Table 1: Descriptive statistics for the tests used in the study

	Partial Dictation	Cloze	Reading	C-test
Mean	6.26	2.95	4.24	32.05
Median	6.00	3.00	4.00	28.00
Mode	3.00	2.00	4.00	17.00
Std. Deviation	4.01	1.85	2.38	15.80
Variance	16.09	3.44	5.68	249.81
Range	16.00	7.00	10.00	65.00
Minimum	.00	.00	.00	4.00
Maximum	16.00	7.00	10.00	69.00

The Cronbach's alpha reliability of the partial dictation test, the cloze test, the C-Test, and the reading comprehension test were .81, .59, .86, and .60, respectively. To compute the reliability of the C-test each passage was considered a super-item or testlet (Eckes & Baghaei, 2015). The reliability of the cloze test and the reading comprehension test are rather small which is due to their small number of items.

4.1 *Correlational Analysis*

Table 2 depicts the correlations between the partial dictation test and the criterion measures and also the correlations between the criterion tests.

Table 2: Matrix of correlations between the tests

	Partial Dictation	Cloze	Reading	C-test
Partial Dictation	1	.57** (.82)	.47** (.68)	.66** (.79)
Cloze		1	.48** (.81)	.56** (.78)
Reading			1	.52** (.73)

** Correlation is significant at the 0.01 level (2-tailed).
Disattenuated correlations are in brackets.

As Table 2 shows performance on the partial dictation is strongly and significantly correlated with the cloze test (r=.57, p<.01, n=112), the C-Test (r=.66, p<.01, n=112), and the reading comprehension test (r=.47, p<.01, n=112). The lower correlation of the dictation with the cloze and the reading comprehension tests could be due to the low reliability of these two criterion measures. Therefore, the disattenuated correlations (correlations corrected for

unreliability) are also reported in Table 2 as well. The disattenuated values indicate that the correlations would have been much higher if the tests were reliable.

Table 2 clearly indicates that there is a positive significant correlation between partial dictation and the reading comprehension test (r=.47, p<.01). The reliability of the reading comprehension test was .60. Low reliability can depress the correlation. Therefore, the corrected correlation for low reliability was computed which was .68. Partial dictation test correlates at (r=.57, p<.01) with the multiple choice cloze test. Therefore, performance on the partial dictation test is strongly related to performance on multiple choice cloze test. The corrected correlation turned out to be .82. Partial dictation correlates at (r=.66, p<.01) with the C-test. The corrected correlation was .79. Therefore, these two measures are highly related. The relatively high correlations found between partial dictation and the criterion tests support the validity of partial dictation as an overall test of general language proficiency.

4.2 Principal components analysis

Principal components analysis was also employed to demonstrate the construct validity of the partial dictation test. The two partial dictation passages, the four C-test passages, the multiple-choice cloze passage and the reading comprehension passage were subjected to principal component analysis (PCA) using SPSS version 21. Prior to performing PCA, the suitability of data for factor analysis was assessed. The Kaiser-Meyer-Olkin value was .87, exceeding the recommended value of 0.60 and Bartlett's test of Sphericity reached statistical significance (p<.001), supporting the factorability of the correlation matrix. Principal components analysis showed the presence of one component with eigenvalue exceeding 1, explaining 58% of the variance. The inspection of the scree plot in Figure 1 revealed one clear break after component one which supports extraction of one factor.

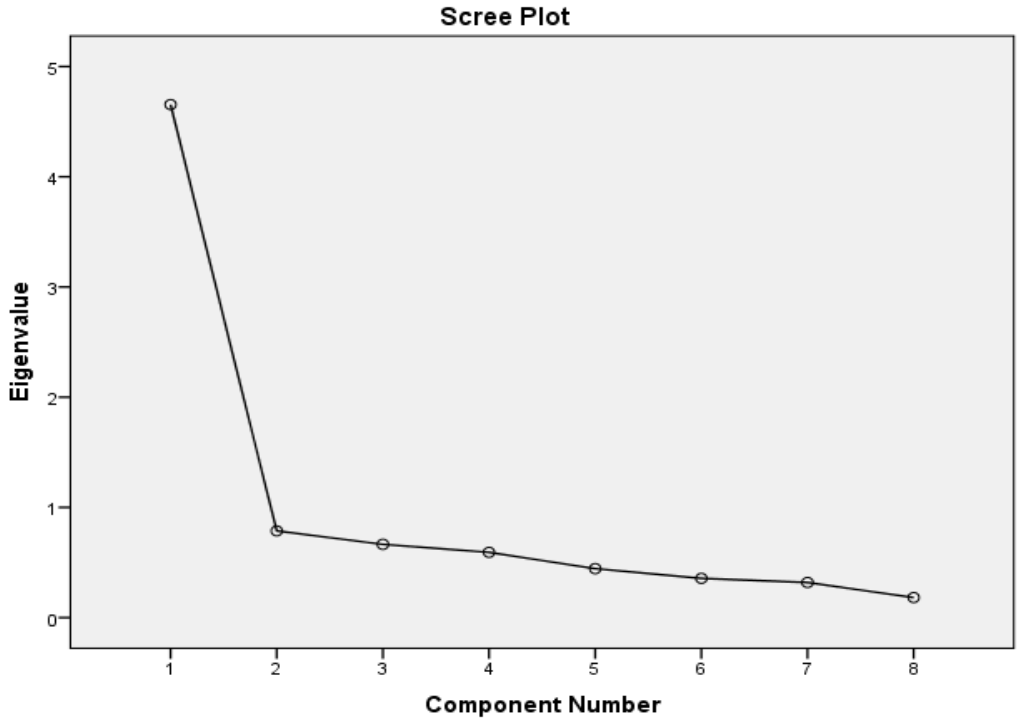

Figure 1: Scree plot for the components

Table 3 shows the loadings of each variable on the single factor extracted from the data. All the loadings are very high (above .66) which supports the unidimensionality of the data. The fact that the partial dictation passages highly load on a single factor on which all other measures load supports the validity of the partial dictation as a measure of overall language ability and reading comprehension.

Table 3: Component matrix and factor loadings

Variable	Component 1
Dictation 1	.787
Dictation 2	.761
Cloze	.716
C-test 1	.758
C-test 2	.824
C-test 3	.829
C-test 4	.749
Reading	.664

5. Discussion and Conclusion

The aim of this study was to examine the validity of partial dictation as a measure of overall language proficiency in English as a foreign language among Iranian EFL learners. One hundred and twelve BA students of English were selected using convenience sampling. Two partial dictation passages along with a multiple-choice cloze passage, four C-Test passages, and a reading comprehension passage containing 12 multiple-choice items were administered to the participants.

The correlations between partial dictation and the criterion measures were strong and significant. If we consider cloze test and C-Test as valid measures of overall language proficiency, the high correlations observed between partial dictation and these measures support the validity of partial dictation as a measure of foreign language general proficiency. Furthermore, the partial dictation had a strong correlation with the reading comprehension test which is evidence that the partial dictation is also a valid measure of reading comprehension.

The criterion measures were also moderately correlated. Zare and Boori (2018) demonstrated a correlation of .80 between cloze test and the C-Test while a correlation of .56 (corrected .78) was observed in this study. They also found strong correlations between C-Test and reading comprehension (r=.83, p<.01, n=150) and cloze test and the reading comprehension (r=.81, p<.01, n=150) while in the present study the correlations between reading comprehension and C-test was .52 (corrected .73) and the correlation between reading and cloze was .48 (corrected .81). The lower correlations observed in this study could be due to smaller reliabilities, smaller sample size, and sampling error.

The correlation between the C-test and reading comprehension was slightly higher than the correlation of reading and the other two tests. This findings suggests that the C-Test is a better measure of reading comprehension than cloze and partial dictation. Although partial dictation has an auditory component its correlations with the C-Test and the cloze test

was higher than its correlation with the reading comprehension test. This is probably because partial dictation, cloze test, and C-test are all members of the family of reduced redundancy tests and have reduced redundancy as their common feature which makes them correlated.

Principal components analysis showed that all the variables, i.e., partial dictation passages, the cloze passage, C-test passages, and the reading comprehension passage, loaded on a single factor with strong loadings. This means that all the variables measure a single dimension. Cloze test and the C-test are established measures of general language proficiency. Since the partial dictation passages strongly load on the single factor where C-tests and the cloze test load the validity of the partial dictation as a measure of general language proficiency is also established.

Developing test items is a challenging task for language teachers. Many teachers do not receive sufficient training in test development and the items that they develop are of poor quality (Alderson, 2005). Many teachers, when in need of items, resort to 'cut and paste' strategy and find suitable items from published test materials available. They usually find their test items among published tests, past test papers, or from textbook and workbook exercises (Coniam, 2009). Furthermore, meeting deadlines for writing items when teachers have many responsibilities is very hard. Therefore, selecting items from existing test items, provided that their content and difficulty is appropriate for the students can guarantee test quality. Coniam (2009) argues that this procedure is justified.

The argument above shows that there is a need for valid tests that are easy to develop. Writing elaborate test items is hard for many teachers who do not have the required expertise in item writing or do not have the time to produce them. This reveals the need for simple and valid tests that can easily be written by every foreign language teacher (Cai, 2012).

Tests of the family of reduced redundancy principle have proven to be easy to construct, administer, and objectively score. The findings of the present study showed that partial dictation tests can easily be constructed by teachers to serve as valid overall language ability tests. The test can be used as a quick and economical instrument that can be used for placement purposes. With careful selection of the texts, syllabus-based partial dictation tests can also be developed for achievement testing.

Although partial dictation was originally suggested as a test of general language proficiency, recently Cai (2012) validated it as a test of listening comprehension. Foreign language teachers can construct partial dictation without much problem to test the listening ability of their learners. Syllabus designers and text book writers can use this test type to measures the overall proficiency of the learners at the end of a course or at beginning to place them in the right level of language learning. The test can also be used to measure progress over time.

Norris (2006) developed and validated a curriculum-based C-Test for university students of German as a foreign language for placement purposes. He carefully selected texts which well represented the expected textual abilities of the candidates and empiricaally demonstrated that C-tests can distinguish learners of various abilities and reliabily place them at different curricular levels. By the same token, all RRP tests can be used for achievement purposes by meticulously selecting texts which match the content of the book learners have covered or the content they are expected to have mastered by the syllabus. In other words, texts should represent what the learners are expected to have learned at a certain ability level to process them.

The partial dictation can also be used as a useful instructional technique as practicing listening comprehension exercises. Ambiguities and noise are quite normal when we listen to announcements in the airports or train stations or when we speak in busy places. Teaching

students to focus on an orally presented texts and listen in for specific missing parts can be a great exercise to train the ears to deal with regular ambiguities in listening tasks in the real world. Therefore, syllabus designers and text book writers can include partial dictation as a listening activity in their material.

5.1 Limitations and desiderata for further research

There are some limitations to this study that can be researched in future.
The researcher only used cloze test, C-test, and reading comprehension as criterion measures in this study. Further studies are needed with more direct tests of language skills as criteria to evaluate the validity of partial dictation. Besides, only correlational methods and exploratory factor analysis was used to gather validity evidence for partial dictation. A comprehensive measure of genral language proficiency such as the TOEFL (Test of English as a Foreign Language) or IELTS (Internatonla English Language Testing System) as a criterion can provide a better picture of the validity of partial dictation. Using grades in the different subjects within the BA English programme such as writing, converstaion, reading, etc. as criteria for validity is also suggested.

Other kinds of validity evidence including fit to latent trait models should also be sought in future (Baghaei & Tabatabaee-Yazdi, 2016; Baghaei, & Shoahosseini, 2019). Item Response Theory (IRT) models have been used to model C-Test (Forthmann, Grotjahn, Doebler, & Baghaei, 2019; Eckes, 2006). The applications of such models to partial dictation tests with their specific dependent structure is a challenge in need of attention. Modeling local item dependence in partial dictation is another IRT-related issue which awaits exploration (Eckes & Baghaei, 2015; Baghaei, & Aryadoust, 2015; Baghaei, 2010b). Investigation differential item functioning and invariance is another issue that should be examined in partial dictation (Baghaei, Bensch, & Ziegler, 2016; Baghaei, 2013; Baghaei, Kemper, Reichert, & Greif, 2019; Ravand, Baghaei, & Doebler, 2019).

We did not consider different variations of the partial dictation as suggested by Cai (2012). Cai (2012) exprimnetd with these variations as listeing tests by examining their factor structure within a set of four different types of partial dictation tests. Investigating the relatiohship between these variations and other tests language skills and other tests of reduced redundency principle can be very informative.

In the standard procedure for developing partial dictation, Johansson (1973) stated that only important and difficult parts of the text should be removed. However, in this study we used random deletion as is common in RRP testing. Whether partial dictation tests with random deletions and those in which "hard and important" parts are deleted are equivalent is a matter of empirical research.

References

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

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

Babaii, E., & Shahri, S. (2010). Psychometric rivalry: The C-test and the cloze test interacting with test takers' characteristics. In Rüdiger Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 41-56). Frankfurt am Main: Lang.

Babaii, E., Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*, 209–219.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal, 9*, 168–175.

Baghaei, P. (2011). *C-Test construct validation: A Rasch modeling approach*. Saarbrücken: VDM Verlag Dr Müller.

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*, 85-104.

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

Baghaei, P., & Grotjahn, R. (2014). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 163-171.). Frankfurt/M.: Lang.

Baghaei, P. (2014). Development and validation of a C-Test in Persian. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp.299-312). Frankfurt/M.: Lang.

Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional Rasch model. *International Journal of Testing, 15*, 71–87.

Baghaei, P. (2010a). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.). *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp.100-112). Frankfurt/M.: Lang.

Baghaei, P., Bensch, D., & Ziegler, M. (2016). Measurement invariance across gender and major in the University of Tehran English Proficiency Test. In Aryadoust, V., & Fox, J. (Eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim* (pp.167-183). New Castle, England: Cambridge Scholars.

Baghaei, P., Kemper, C., Reichert, M., & Greif, S. (2019). Mixed Rasch modeling in assessing reading comprehension. In Aryadoust, V. & Raquel, M. (Eds.), *Quantitative Data Analysis for Language Assessment* (Vol. II) (pp. 15-32). New York: Routledge.

Baghaei, P. (2010b). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling, 52*, 313-323.

Baghaei, P., & Shoahosseini, R. (2019). A note on the Rasch model and the instrument-based account of validity. *Rasch Measurement Transactions, 32*, 1705-1708.

Baghaei, P., & Ravand, H. (2019). Method bias in cloze tests as reading comprehension measures. *Sage Open, 9*, 1-8. doi: 10.1177/2158244019832706

Broukal, M. (2002). *TOEFL reading flash*. Connecticut, USA: Peterson's.

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

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the impact of training in test development principles on improving test quality. *System, 37*, 226–242.

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependency in C-Tests. *Applied Measurement in Education, 28,* 85–98.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*, 290-325.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 1-44). Frankfurt am Main: Lang.

Fadaeipour, A., & Zohoorian, Z. (2017). Comparing the psychometric characteristics of speeded and standard C-Tests. *International Journal of Language Testing, 7*, 40-50.

Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2019). A comparison of different item response theory models for scaling speeded C-Tests. *Journal of Psychoeducational Assessment.* Advance online publication. Doi: 10.1177/0734282919889262

Gaies, St. J., Gradman, H.L., & Spolsky, B. (1977). Toward the measurement of functional proficiency. *TESOL Quarterly*, *11*, 51-57.

Grotjahn, R., Schlak, T., & Aguado, K. (2010). S-C-Tests: Messung automatisierter sprachlicher Kompetenzen anhand von C-Tests mit massiver textspezifischer Zeitlimitierung. In Rüdiger Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 297-319). Frankfurt am Main: Lang.

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

Holster, T. (2017). Cloze-elide as a classroom test. *Shiken, 21*, 1-18.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

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.

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

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing, 1*, 134-146.

Manning, W. H. (1987). Development of cloze-elide tests of English as a second language. *ETS Research Report Series*. Princeton, NJ: Educational Testing Service.

Norris, J. M. (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. In Grotjahn, R. (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 45-83). Frankfurt am Main: Lang.

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.

Ravand, H., Baghaei, P., & Doebler, P. (2019). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology*. Doi: 10.3389/fpsyg.2019.02930

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

Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing, 8*, 12-26.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt/am: Peter Lang.

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

Spolsky, B. (1968). *What does it mean to know a language? Or how do you get someone to perform his competence?* Paper presented at the Second Conference on Problems in Foreign Language Testing, University of Southern California.

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.

Zare, S., & Boori, A. A. (2018). Psychometric evaluation of the speeded cloze-elide test as a general test of proficiency in English as a foreign language. *International Journal of Language Testing, 8*, 33-43.