LUBLIN STUDIES IN MODERN LANGUAGES AND LITERATURE,
43(3), 2019, HTTP://LSMLL.JOURNALS.UMCS.PL

Ferit Kılıçkaya
Burdur Mehmet Akif Ersoy University, Turkey
ferit.kilickaya@gmail.com
https://orcid.org/0000-0002-3534-0924

Assessing L2 vocabulary through multiple-choice, matching, gap-fill, and word formation items

ABSTRACT
The current study aims to determine the effect of multiple-choice, matching, gap-fill and word formation items used in assessing L2 vocabulary on learners' performance and to obtain the learners' views regarding the use of these types of items in vocabulary assessment. The convenience sampling method was selected, and the participants of the study included 30 freshmen enrolled in the General English course offered in the Department of Public Administration at a state university in Turkey. The main findings revealed that the participants considered the multiple-choice and matching items were easy to understand and to answer and that gap-fill and word formation items were difficult due to several reasons.
Keywords: assessing vocabulary, multiple-choice, matching, gap-fill,, word formation

1. Introduction

Vocabulary assessment is an indispensable aspect of language teaching as Nation (2008) clearly indicates the aim of vocabulary assessment is "to work out what needs to be taught, to monitor and encourage learning, to place learners in the right class, to measure learners' achievement by giving a grade, and to measure learners' vocabulary size or proficiency" (p. 144). Teachers need to determine to what extent the words that they taught have been mastered by the students both receptively and productively. However, like assessing other language skills and components, L2 vocabulary assessment poses a challenging task for language teachers due to several reasons (Shen, 2003). One is that the assessment format, technique or task used in the assessment practices may have a beneficial or harmful effect on learners' performance. Another reason is that preparing appropriate items for different formats for vocabulary assessment requires knowledge and expertise since each might have several advantages and disadvantages.

DOI: 10.17951/lsmll.2019.43.3.155-166

One of the major issues for learners regarding L2 vocabulary is producing the words in addition to recognizing it (McCarthy, O'Keeffe, & Walsh, 2010). While recognizing words includes differentiating words from others and recalling the meaning, producing the words might pose serious issues since it includes not only forming and writing words but also recalling the meaning. In order to overcome these issues, in teacher education programs and in-service language teachers are presented, taught and asked to practice several formats or techniques to assess vocabulary receptively and productively (Ur, 2012). Moreover, in-service training programs also include these formats or techniques to keep the in-service teachers up to date with vocabulary assessment. Of these, multiple-choice (MC), matching, gap-fill and word formation (WF) formats (Heaton, 1990; Hughes, 2003; Brown, 2005; Brown & Abeywickrama, 2010; Bailey & Curtis, 2015) are among the most commonly used items in the language classrooms, in the nation-wide and worldwide conducted exams such as Cambridge English: First (FCE).

2. Formats to Assess vocabulary

The formats to assess vocabulary can be divided into two kinds: recognition based items and productive based items (Heaton, 1990; Hughes, 2003; Brown, 2005; Riahi, 2018). Recognition based or oriented items include the most common items such as MC and matching, while productive ones include items such as gap-fill and cloze tests (Brown & Abeywickrama, 2010; Bailey & Curtis, 2015; Brown & Trace, 2017). These types of formats are considered much more challenging and demanding for the learners as they have to consider the meaning of the word and to provide the correct form (Read, 2012). The formats used in the current study will be briefly discussed below, indicating the main features of them.

MC format

MC items were, and still are, one of the most common formats used in language tests, mainly used to assess grammar and grammar. MC items include a statement, called as the stem, which a question to be answered, a problem to be solved, or as in most situation, an incomplete statement to be completed, and the response options to be used in the blank in the stem (Bailey & Curtis, 2015). Of the options, the correct one is considered as the key (correct) answer, while the others are called distractors, which are the incorrect responses that are used to distract the responders that who do not know the correct answer. The total number of options including the correct answer ranges from four to five depending on the needs and the level of the students. In high-stakes exams, incorrect answers (generally four) provided may cancel out one correct answer in order to refrain the test-takers from benefiting from their guessing skills. There are several advantages of using MC items. One is that scoring the answers is relatively easier and practical compared to other formats that aim to assess vocabulary, and it is more objective in terms

of scoring (Bachman & Palmer, 1996). However, creating good MC items is not easy, as it requires expertise and experience in producing well-structured items (Read, 2000). Another disadvantage of MC items is that they cannot be used to assess productive skills. In other words, assessment will be based on recognition of the correct answer, rather than producing it.

Matching format

Matching items are as popular MC items, and they are generally used in assessing vocabulary. The basic format of matching items includes two columns of information. The left column includes the explanations, statements or the definitions of the words. The right column, on the other hand, includes the words or the options. Learners are then asked to match the words/options on the right with the words/statements on the left by generally writing the letters (A, B, C, D…) that correspond to the options on the right column. One of the main advantages of using this format it that more distractors can be provided (Miller, Linn, & Gronlund, 2013). While, for example, in MC items, 3 or 3 distractors are possible, in matching format, there might be 10 or even more. However, this format is still based on recognition, rather than the production of the correct answer.

Gap-fill format

Unlike MC and matching formats, gap-fill format allows creating items that encourage learners to produce vocabulary. In the gap-fill format, learners are provided with sentences that have gaps. Learners are expected to read each sentence and to provide the suitable word that may complete the sentence. In other words, learners have to produce the word rather than just recognize it. From this perspective, gap-fill format provides teachers the opportunity to construct questions that assess learners' production of vocabulary. Constructing gap-fill items is rather easier compared to other formats. However, several disadvantages can also be associated with this format. One of these disadvantages is that students' producing the answer in order to complete gap requires more time compared to MC and matching items produce the answer in order to complete the gap (Coombe, Folse, & Hubley, 2007). Another disadvantage is that learners might come up with possible answers although they might not be the one in the teacher's mind or key to the test.

WF items

WF items are mainly used for assessing lexical knowledge; however, structural knowledge might also be required. In high-stakes exam such as Cambridge English Proficiency, the focus of this format is "on vocabulary, in particular, the use of affixation, internal changes and compounding in WF" (Cambridge English Proficiency, 2016, p. 7). In this format, several words are taken from a text and

the stem words are provided at the end of the lines as separated from the text. The learners are then asked to complete each gap using the appropriate form of the word given as the stem word. Learners are expected to use affixes, internal changes, and compounds while forming the words based on the stem. However, it is also required to consider the context in which each gap is provided since learners are to provide the appropriate part of speech such as the noun, adjective, or adverb form of the stem provided.

3. Related research on vocabulary assessment

A plethora of research has been conducted on teaching L2 vocabulary, and the research conducted has yielded varying results. Moreover, the most common techniques and formats have started to be used with the new advances in technology, resulting in promising learning gains (Özer & Koçoğlu, 2017). However, there are few studies conducted on the use of different assessment forms in vocabulary assessment and the learners' views. It is not rare to observe that language teachers placing less importance on vocabulary assessment, if not totally ignoring it and assessing learners' vocabulary knowledge only asking students to provide meanings in their L1 (e.g., Tuyen, 2015).

The study conducted by Amini and Ibrahim-González (2012) investigated the effects of cloze and MC tests on the thirty freshmen students majoring in English language teaching at a university in Iran. The results indicated that teaching and testing vocabulary through cloze tests encouraged students to use the vocabulary productively rather than receptively since the participants tried to infer the meaning benefiting from the context provided. Another study by Kremmel and Schmitt (2016) investigated whether the results of assessments that included various item formats could provide information on the learners' ability to use words. In other words, the study tried to indicate whether the participants, having provided correct answers on the vocabulary test, could use those words in other situations that required other skills such as reading. The participants included eighteen English native speakers and twelve non-native English speakers at a School of English at a British university and responded to vocabulary questions in four different item formats (multiple matching, MC, and two types of cloze). The results indicated that these four item formats might not indicate whether the correctly answered items could be employed by the participants in reading. Therefore, it was put forward that the scores obtained through these formats could not be used to go beyond the form-meaning link.

The study conducted on the use of gap-fill (Kılıçkaya, 2011) compared the participants' performance on the same time items that were presented in different forms. The study included three groups. The participants in the control group were presented with a paragraph with blanks and asked to select the best option to fill the gaps. However, the participants in the first experimental group were asked to select the best option on the individual sentences taken from the same paragraph,

while the ones in the second experimental group were asked to fill in the blanks in the same paragraph with no options to select. The results indicated that the participants in the first experimental group outperformed the others. The results also showed that the participants in the second group obtained the lowest scores, as they were not able to provide the words although they guessed what would come to the gap considering the meaning.

4. The current study

The current study aimed to determine the effect of several vocabulary assessment formats (MC, matching, gap-fill, and WF) in assessing L2 vocabulary on learners' performance. The study also aimed to obtain the participants'' views regarding the use of these formats in vocabulary assessment in the classroom. In line with these purposes, the following research questions were put forward:

1. What is the effect of using different formats in assessing L2 vocabulary in learners' performance in the exercises?
2. What are the participants' views on these formats?

5. Methodology

Research design

The study adopted a mixed-method approach by utilizing both quantitative and qualitative data. The quantitative data included the participants' scores on different formats using in vocabulary assessment. The qualitative data included the participants' responses obtained during the semi-structured interviews.

Participants

The participants of the study were 30 freshmen enrolled in the General English course offered in the Department of Public Administration at a state university in Turkey. Of the participants, 14 were female, while 16 were male. The participants' age ranged from 18 to 22, with an average of 19.2. Most of the participants were graduates of high schools, while only 4 of them were a graduate of vocational schools with 2-year education. The participants were enrolled in the General English course, which aims to have learners learn the basic grammatical structures and to produce sentences that will achieve communicative functions in both written and spoken English. The course was offered three hours each week for fourteen weeks, with 42 hours in total.

Data collection instruments

The data collection instruments included twenty-vocabulary assessment activities that included MC, matching, gap-fill and WF items and the semi-structured interviews. After each unit, the participants were provided four exercises, each of which included different assessment items. There were 5 items in each exercise,

and these items were the frequently used words in the student book, the workbook, and the supplementary materials provided by the lecturer in the classroom. The example items used in the study are presented in the Appendix. These items are based on the content of the book Face2face: Elementary student's book, written by Redston and Cunningham (2012). Semi-structured interviews were conducted at the end of the seventh week with randomly selected ten students regarding their performance in the exercises as well as their views towards the use of different assessment items in the vocabulary exercises. The interviews took place in the researcher's office in the participants' native language (Turkish) and took 7.5 minutes on average.

Data collection procedure

During the first week, the participants were introduced to the course and then asked if they would like to participate in the study in which different assessment formats would be used to assess vocabulary. No further details were provided regarding the study. After obtaining their consent, they were informed that at the end of each unit (the first five units), there would be assessment exercises in different formats. They were also informed that these would not affect their final grades in the course and the results would be just used for the analysis of the effects of using different formats in exercises. Then, after each unit, the participants were asked to do the exercises in four different exercises, each of which included five items. The total number of items in each session was 20 and these items included the most frequently used words in the book as well as the supplementary materials. At the end of the sixth week, the participants completed the last exercises, which were followed by the semi-structured interviews. These interviews were conducted with randomly selected ten participants at the end of the seventh week just before the midterms.

Data analysis

The quantitative data obtained from the participants' exam results in different formats were subject to statistical analysis using IBM SPSS 24. One-way repeated measures ANOVA was conducted to determine any statistically significant differences between the means of these four formats, namely, MC, matching, gap-fill and WF. Moreover, the qualitative data obtained through the semi-structured interviews on the participants' views regarding these formats were transcribed verbatim. Later, the transcriptions were subject to content analysis to determine the emerging themes and codes. Several responses were selected as the quotations and were translated into English.

6. Findings and Discussion

The quantitative and qualitative findings will be presented together in this section since the results are related to each other. A repeated measures ANOVA with a Greenhouse-Geisser correction determined the average scores obtained differed statistically significantly among the item formats ($F(2.720, 78.893) = 34.062$, $p < 0.05$). A statistically significant difference existed among the four sets of scores. The effect size calculated as multivariate partial eta squared was determined to be = .99, which suggests a very large effect size. The participants obtained the highest average on the questions in the MC format ($\bar{X}$ =15.73) and the lowest on the questions in the WF format ($\bar{X}$=10.30).

The pairwise comparisons were also conducted to determine which set of scores obtained on different types of cloze procedure differed from one another. Post hoc tests using the Bonferroni correction were also conducted. The summary of the results is provided in Table 1. These tests revealed that the participants' scores obtained on the MC and matching tests differed significantly from the gap-fill and WF ones, with the difference found to be significant at the 0.05 level. However, no statistically significant difference was found between the two item formats: MC and matching. This finding clearly indicates that the item formats, such as MC and matching, remain popular among learners due to their apparent aptness for testing vocabulary. The great majority of the participants (n=9) indicated during the interviews that compared to other item formats, MC and matching items were relatively easier as they did not have to provide the form or the meaning but 'recognize' the best word that would complete the blank.

One of the participants expressed this clearly as follows:

> *Providing the suitable words for the gaps was difficult. I was required to remember the form, I mean, the spelling of the word. Similarly, WF was also challenging. However, when it comes to MC or matching questions, I was rather relaxed, as I did not have to produce but select the best word* (Male, ID 8).

This finding is consistent with that of the study conducted by Kılıçkaya (2011), indicating that MC and matching item formats led the participants to better use of receptive knowledge rather than the productive one and with that of the discussion on the challenges imposed by productive formats (Read, 2012). Task or item familiarity is known to affect the results and the performance of the candidates in addition to the reliability of the exams conducted (Brown & Abeywickrama, 2010). This might be attributed to the fact that the participants' receptive vocabulary can be much larger than the productive one (Coxhead, 2018), and therefore, it might lead them to perform better in the MC and gap-fill than the WF. This might also be because these participants were much more familiar with these activities both in the classroom and outside the classroom, which requires caution while considering the effects.

The statistical analysis also indicated that the participants obtained the lowest

Table 1. *The Pairwise Comparisons among the average scores obtained from item formats*

(I) Item format	(J) Item format	Mean Difference (I-J)	Std. Error	Sig.
MC	Matching	-.533	.819	1.000
	Gap-fill	4.733*	.612	.000
	WF	5.433*	.733	.000
Matching	MC	.533	.819	1.000
	Gap-fill	5.267*	.776	.000
	WF	5.967*	.873	.000
Gap-fill	MC	-4.733*	.612	.000
	Matching	-5.267*	.776	.000
	WF	.700	.678	1.000
WF	MC	-5.433*	.733	.000
	Matching	-5.967*	.873	.000
	Gap-fill	-.700	.678	1.000
*. The mean difference is significant at the .05 level.				

scores in the questions created in gap-fill and WF item formats. The participants' average score was 11.00 for the gap-fill questions and 10.3 for the WF items. Although there was no statistically significant difference between the scores in gap-fill and WF items, the participants stated that WF items were rather difficult and provided several reasons for this. The common reasons stated were determined to be related to the characteristics of the item and the knowledge required to provide the correct answer (n=7). One of the participants explained this as follows:

> *These types of activities [gap-fill and WF] required writing the answers instead of selecting the correct answer. Compared to other item formats, especially WF was, I think, difficult. The reason is that it was testing also the word structure [part of speech] and some structures [syntactical knowledge]* (Female, ID 4).

This type of activity aimed to require the participants to demonstrate their understanding of the meaning considering the context, as the participants agreed, it was shown to be testing syntactical knowledge (Stopar, 2014). That is, without knowing much about the meaning of the word to be inserted into the blank, the participants tried to determine whether it would be an adjective, a verb, or a noun. The results suggest that the majority of the participants complained that WF items required the knowledge of syntax and morphology since through this knowledge it was possible to determine the correct part of speech. Moreover, since the participants were not used to be assessed through productive knowledge (Toksöz & Kılıçkaya, 2017), it was possible that they found gap-fill and WF more

challenging compared to other item formats.

As indicated by Schmitt and McCarthy (1997) and McCarthy (2003), linguistic contexts especially aid the learners' ability of utilizing morhphonological as well as lexical rules, which facilitates understanding of the meaning and the form of the word. Therefore, some participants also valued the use of contexts. In other words, rather than just asking the meaning of a word given in isolated forms without using it in a sentence was highly valued by the participants.

Considering the findings obtained, it can be stated from the pedagogical perspective that Failing to encourage learners to produce the word (pronunciation) and write it (spelling) would be tantamount to dereliction of the basic duty in teaching and learning any foreign language, not just English (Milton & Hopkins, 2006). Therefore, it is suggested that teachers should introduce productive tasks and items into the classroom such as WF in addition to the common exercises. One suggested activity can be that learners might be asked to use the common words written on paper through using (speaking) them in context.

7. Conclusion and suggestions for further research

The current study aimed to determine the effects of using various items in assessing L2 (English) vocabulary on the university students' performance in these items and the students' views on the use of different items. The study used both quantitative and qualitative data to achieve this aim. The results mainly indicated that gap-fill and WF items were found to be rather difficult by the participants due to several reasons. These items required the participants to produce the words based on several factors such as the context, meaning, and the part of speech. Therefore, most participants found them more demanding compared to other times. These findings were also confirmed by the quantitative findings regarding the participants' performance on the tests. The findings also point towards the need for more use of productive assessment, instead of recognition assessment in testing learners' lexical knowledge. The need for this is also reflected in the participants' responses during the interviews.

The quantitative data collection instrument in this study focused on the written form of the words. Therefore, further research can also use other item formats to test learners' phonological as well as orthographic vocabulary knowledge and determine the effects of the assessment of these types of knowledge on the participants' performance.

Acknowledgment. This article is the revised and extended version of the oral paper presented at the 7th International Conference on Narrative & Language Studies (2018), in Trabzon, Turkey.

References

Amini, M., & Ibrahim-González, N. (2012). The washback effect of cloze and multiple-choice tests on vocabulary acquisition. *Language in India, 12*(7), 71-91. Retrieved March 8, 2018, from http://www.languageinindia.com/july2012/v12i7july2012.pdf.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, K. M., & Curtis, A. (2015). *Learning about language assessment: Dilemmas, decisions, and directions* (2nd ed.). Boston, MA: National Geographic Learning.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). New York, NY: Pearson.

Brown, J. D. (2005). *Testing in language programs*. New York, NY: McGraw-Hill.

Brown, J. D., & Trace, J. (2017). Fifteen ways to improve classroom assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 3, pp. 490-505). New York, NY: Routledge.

Cambridge English Proficiency. (2016). *Cambridge English proficiency: Handbook for teachers for exams from 2016*. Cambridge: Cambridge English Language Assessment. Retrieved March 8, 2018, from http://www.cambridgeenglish.org/images/168194-cambridge-english-proficiency-teachers-handbook.pdf.

Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: The University of Michigan Press.

Coxhead, A. (2018). Vocabulary assessment. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-6). John Wiley & Sons, Inc. DOI: 10.1002/9781118784235.eelt0628.

Heaton, J. B. (1990). *Writing English language tests*. New York, NY: Longman.

Hughes, A. (2003). *Testing for language* teachers (2nd ed.). Cambridge: Cambridge University Press.

Kılıçkaya, F. (2011). Yabancı dil sınavlarında boşluk tamamlama sorularının öğrenci başarısındaki etkisinin karşılaştırılması [Comparing the effects of gap-fill on the test takers' performance in the foreign language tests]. *International Conference on New Trends in Education and Their Implications Papers* (pp. 80-86). Ankara: Siyasal Kitabevi. Retrieved March 8, 2018, from http://www.iconte.org/FileUpload/ks59689/File/014..pdf.

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly, 13*(4), 377-392. DOI: 10.1080/15434303.2016.1237516

McCarthy, M. (2003). *Vocabulary*. Oxford: Oxford University Press.

McCarthy, M., O'Keeffe, A., & Walsh, S. (2010). *Vocabulary matrix: Understanding, learning, teaching*. Hampshire: Heinle, CENGAGE Learning.

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). *Measurement and assessment in teaching* (11th ed.). New York, NY: Pearson.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. Canadian Modern Language Review, *63*(1), 127-147. Retrieved June 10, 2018, from https://utpjournals.press/doi/pdf/10.3138/cmlr.63.1.127.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle Cengage Learning.

Özer, Y. E., & Koçoğlu, Z. (2017). The use of Quizlet flashcard software and its effects on vocabulary learning. *Language Journal, 168*(1), 61-81. Retrieved March 8, 2018, from http://dergiler.ankara.edu.tr/dergiler/27/2188/22676.pdf.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2012). Assessing vocabulary. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 257-264). Cambridge, MA: Cambridge University Press.

Redston C., & Cunningham, Gillie. (2012). *Face2face: Elementary student's book* (2nd ed.). Cambridge, MA: Cambridge University Press.

Riahi, I. (2018). Techniques in teaching and testing vocabulary for learners of English in an EFL context. In S. Hidri (Ed.), *Revisiting the assessment of second language abilities: From theory to practice* (pp. 289-310). Cham, Switzerland: Springer International Publishing.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge, MA: Cambridge University Press.

Shen, W.-W. (2003). Current trends of vocabulary teaching and learning strategies for EFL setting. *Feng Chia Journal of Humanities and Social Sciences, 7,* 187-224. Retrieved June 10, 2018, from http://www.fcu.edu.tw/wSite/public/Attachment/f1378105968860.pdf.

Stopar, A. (2014). Testing and assessing English word-formation. *Romanian Journal of English Studies, 11*(1), 1-8. DOI: 10.2478/rjes-2014-0033.

Toksöz, İ., & Kılıçkaya, F. (2017). Review of journal articles on washback in language testing in Turkey (2010-2017). *Lublin Studies in Modern Languages & Literature, 41*(2), 184-204. DOI: 10.17951/lsmll.2017.41.2.184.

Tuyen, L. V. (2015). An investigation into the effectiveness of learning assessment for non-English major students at the tertiary level. *International Journal on Studies in English Language and Literature, 3*(5), 14-31. Retrieved June 10, 2018, from https://www.arcjournals.org/pdfs/ijsell/v3-i5/3.pdf.

Ur, P. (2012). *A course in English language teaching* (2nd ed.). Cambridge, MA: Cambridge University Press.

APPENDIX - Example Items

MC items

(1) My -----‘s name is Ahmet and we’ve got two children.
A) husband B) wife
C) sister D) father

(2) I ----- work at 7.00 a.m. every morning.
A) have B) go
C) sleep D) start

Matching items

1. My sister is a/an -----. She tries to prevent crime. 2. Nejat İşler is a/an -----. You can see him in the movies. 3. His brother is a/an -----. He helps ill people. 4. My father is a/an -----. He repairs cars.	A) doctor B) lawyer C) actor D) engineer E) mechanic F) police officer

Gap-fill items

(1) This jacket is cheap. It ----- only 5 TL.
(2) How ----- are these t-shirts?
(3) How ----- months are there in a year?
(4) ----- mobile phone is this? It is Mary’s.
(5) I ----- breakfast at about 7.30 in the morning.

WF items

(1) I go ----- every week to keep healthy. (SWIM)
(2) Ayşe is ----- because she can’t come to the party. (HAPPY)
(3) We like the new English Teacher because she is very -----. (FRIEND)
(4) His new car is ----- beautiful. It looks great. (REAL)