# The Effects of Dictionary Use and Pair Work on Cloze Testing Performance in the ESL Classroom

# 英語を母国語としない日本人のための英語クラス（ESL）での穴埋めテスト能力における辞書使用の有無とペアワークの効果

Ronald Kibler[1], Patricia Yamada, Karen Cline, Gregory Samsonow, Jon Cartwright
ロナルド・キブラー[2], パトリシア・山田, カレン・クライン,
グレゴリー・サムソノー，ジョン・カートライト

## 要　旨

本研究の目的は、穴埋めテスト能力における辞書使用とペアワーク（話し合い）の有無の効果を準実験デザイン法によって比較することを目的とした。日本人の大学２年生（N＝110）を対象者として、４つのグループ（①ペアで辞書を使用、②ペアで辞書を使用しない、③1 人で辞書を使用、④１人で辞書を使用しない）に対象者を分け、すべてのグループに同じ穴埋めテストを実施した。穴埋めテスト能力における条件間の比較には二元配置分散分析および多重比較を用いた。なお、統計的有意性は危険率５％未満で有意性ありとした。穴埋めテストは、辞書（p<0.001）とペアワーク（p<0.05）を用いた両方の主効果が認められた。多重比較検定の結果では、①ペアで辞書を使用したグループと③1 人で辞書を使用したグループとの間に有意な差がなかった以外は、すべての組み合わせにおいて有意な差が認められた。以上の結果からペアワーク(話し合い)での穴埋めテストにおいて、高い応用力の活用を容易にすることが示唆された。

## Abstract

In this study the effects of dictionary use and pair work on cloze test performance were compared using a quasi-experimental design. Japanese second year university students (N=110) were assigned to one of the following testing conditions: 1) working in pairs with a dictionary, 2) working in pairs without a dictionary, 3) working alone with a dictionary, 4) working alone without a dictionary. All groups were given the same test. The results of a 2x2 factorial ANOVA revealed that the interaction of dictionary use and discussion was statistically significant, as were the main effect of dictionary use, and the main effect of pair work. Examination of the simple effects revealed that the use of dictionaries did not have as great an effect on cloze test performance as pair work and discussion. One explanation for the higher efficacy of pair work could be that pair work and discussion facilitates the use of higher order thinking skills, resulting in the attainment of higher cloze test scores.

キーワード:辞書使用、討論、ペアワーク、穴埋めテスト、構成主義
*Keywords*: dictionary use, discussion, pair work, cloze testing, constructivism

[1] Liberal Arts Education Center, Sapporo Campus, Tokai University, 5 - 1 - 1 - 1 Minamisawa, Minami - ku, Sapporo005 - 8601, Japan; E-mail: ronkib(a)gmail.com.

[2] 東海大学札幌教養教育センター，005- 8601 札幌市南区南沢 5 条 1 丁目 1-1

# The Effects of Dictionary Use and Pair Work on Cloze Testing Performance in the ESL Classroom

It is not too much to say that our ability to create and use language is no small part of what makes us human. It is something both basic and common to all of us. Yet, when it comes to how we actually learn language, there is much that is still not very well known or understood. The understanding of how people learn a second language is even murkier. That there is so much to consider when we try to understand how people create and use language. And because language learning happens on an individual basis, the variety and uniqueness of learning situations is profound. Trying to understand how people learn languages is more akin to looking through a kaleidoscope, where the viewer is somewhat overwhelmed with the shifting masses of visual information, than staring through a microscope, where focus is drawn down to a few clear and distinct details. However, for all the diversity the puzzle of language learning presents, underlying factors do arise. This report is about two factors that are often addressed in the discussion of language learning, dictionary use and pair work. Pair work is considered to be synonymous with discussion.

## Dictionary use

When it comes to language, a dictionary is considered to be a very important resource. It can be used to check meanings, spellings, pronunciation, grammatical information about words, and it may also give examples of how a word is used or translated. With its concise, consolidated wealth of information, a dictionary can be a powerful language learning aid.

Perhaps it is just this, its excellent quality and quantity of information, as well as its nearly ubiquitous presence in schools, homes, and libraries that can cause it to be seen as a cornerstone of language learning: the gatekeeper, or even a panacea for so many of the pitfalls faced by students in learning a language.

It is not uncommon for teachers in Japan to have a mandatory ‘bring your dictionary to class’ policy. Such teachers strongly feel that without a dictionary, classroom learning is impeded. But when it comes to testing, dictionary use is not always encouraged or allowed. In the United States, the regulations applying to use of dictionaries by English Language Learners on state exams sometimes allow, but sometimes forbid, dictionary use (Albus, Thurlow, Liu, & Bielinski, 2005). There is a tension here between thinking that a dictionary is a 'must have' for language learners, while at the same time seeing it as a threat to testing which would make the answers too obvious.

Upon more careful consideration, it becomes clear that a dictionary is not really either of these. It is what it is, a well organized, fairly accessible reference for many language questions. While it is excellent for referencing information, in terms of language production is not usually essential (Schofield, 1982). It makes sense that stopping to look things up in a dictionary is not easy when it comes to using a foreign language in a number of real life situations. When talking to a friend, asking questions in class, or performing timed reading and writing tasks, the use of a dictionary is not really advantageous, and is not usually essential. If, in the course of one's daily routine, EFL students stop to look up every word they don't know, they will probably spend too much time consulting the dictionary.

In a number of studies, the effect of dictionary use on test performance has not been found to be significant. After a pilot study of 900 EFL students' performance on reading tests, Bensoussan found that, 'There was no significant difference in test scores among those who used the dictionary and those who did not", (Bensoussan, Sim, & Weiss, 1984, p. 265). In 1984 M. Bensoussan, D. Sim, and R. Weiss carried out a landmark study examining the effect of dictionary usage on reading test performance of EFL students at Israeli Universities. The findings came from three studies spanning two years and involving 1,501 participants. In each case dictionary usage had no significant effect on EFL test performance. At the time of this study one of the main issues revolving around dictionary use was the advantages of monolingual, as opposed to bilingual, dictionaries. Although students involved in the study showed a preference for bilingual dictionaries, neither was shown to affect test performance.

In the case of reading comprehension, however, dictionary use was found to have an effect on the vocabulary acquisition of 105 native speakers of English learning university Spanish (Knight, 1994). While this would support the argument for the usefulness of the dictionary as a means of explicit vocabulary acquisition, it does not contradict studies that show a statistically non-significant effect of dictionary use on test taking.

With the advances in dictionary technology the emphasis of dictionary use studies has shifted from the monolingual-bilingual paper dictionary debate to the use of electronic dictionaries, hypertext, and most recently applications (apps) and touch screens. Still, a number of more recent studies continue to suggest that the effects of dictionary use on testing performance are not significant. In a study involving 80 undergraduates in a fifth semester university Spanish classes, Aust, Kelly, and Roby found that students using technologies that accelerated and eased the information search process, such as hyper-references, would look up words at more than double the rates of students using conventional dictionaries. However, measurements of differences in comprehension between conventional dictionary users and hyper-reference users were not significant (Aust, Kelly, & Roby 1993). More recently, in a study on reading test performance of English language learners using an English dictionary, a group of 113 Hmong students in the eighth grade showed that the effect of dictionary accommodation was not significant (Albus, Thurlow, Liu, & Bielinski, 2005).

Although there are many peripheral issues surrounding dictionary use, such as basic familiarity with using dictionaries and dictionary skills, the type of dictionary, level of the student, the nature of the setting in which a dictionary is used (reading versus testing), time constraints and so on, research shows that in a variety of testing situations, dictionary use will not have an effect on test performance.

## Pair Work and Discussion

Assigning students to work in pairs is a common practice in EFL classrooms, and is considered to be an effective way to help students learn for a number of reasons. One obvious advantage is that it gives learners more of an opportunity to practice the L2 in ways that provide both a quality and quantity of interaction that is not present in teacher-fronted classes (Storch & Aldosari, 2013). As opposed to listening and note taking, working in pairs can be considered synonymous with having a discussion. From a cognitive perspective, pair work, when seen as a form of cooperative learning, leads to better overall learning and retention. It has been shown to allow for higher rates of time on task, and quality of reasoning (Johnson & Johnson, 2009).

While pair work has been shown to be effective for teaching and learning, it does require strategies for dealing with individual differences in ability (Stroch & Aldosari, 2013). Assessment of pair work can be problematic. One of the difficulties of pairing and assessment is that the way in which students are paired may affect their performance and evaluation. However it has been shown that differences in proficiency is not an insurmountable obstacle to the use of paired oral assessments (Davis, 2009). While the assessment of cloze test performance in this study is different from the types of oral assessments Davis investigated, there was no reason to suppose that pair work and discussion would not lead to a fair assessment of the participants' performance on the test.

**Cloze testing**

Cloze testing is an 'item deletion' method of testing that was introduced in 1953 by W. L. Taylor as a way to assess 'the relative readability of written material for school children in the United States' (Brown, 2002). It has been a popular way of testing English language learners for decades, but in many ways the cloze test is an enigma. What the cloze does, how it works, and what it is measuring is often difficult to explain, and there can be problems related to reliability and validity if the test is not designed carefully. One might suppose that the creation of an item deletion test would be easy, but in fact it takes a lot of work to come up with a test that will accurately reflect differences in student abilities (Brown, 1998; Chapman 2007). But cloze tests do work. They have been used as a part of the Michigan Examination for the Certificate of Proficiency and were shown to have construct validity (Saito, 2003). Cloze tests have also been shown to correlate significantly with scores on the TOEFL exam (Saeedi, Tavakoli, Kazerooni, & Parvaresh, 2011).

In spite of its difficulties and limitations, the cloze test seemed especially well adapted to this study of dictionary use and pair work. Since the role of vocabulary in the cloze test is especially important, it seemed a good avenue of investigation into dictionary use and testing. On the other hand, the puzzle-like nature of cloze tests seemed a suitable way to accommodate investigation of the effects of pair work and discussion on test performance. Issues related to sampling affected the design of the cloze test used in the study. Random sampling was not possible; participants were not able to gather in the same place, at the same time. This necessitated designing a single cloze test that could be taken by four classes with slightly different, but not significantly different, levels of English skills, and still be counted on to produce test score means that were both valid and reliable. While, in the end, that proved to be achievable, it was not an easy target to hit.

## Assumptions

The assumptions of this study are as follows:

$H_o^1$: There will be no significant interaction between the main effects of dictionary use and pair work (discussion).

$H_o^2$: There will be no significant difference between the test scores of participants using dictionaries and those not using dictionaries on the cloze test

$H_0^3$: There will be no significant difference between the test scores of participants working in pairs and participants working alone on the cloze test.

## Method

## Participants

A total of 320 Japanese university English students participated in the pre-testing phase of the study. After selection, 110 of the selected students took the cloze test. 79 participants were from the Engineering Department, and 31 were from the International Culture Department. The average age of the participants was 19.7 years old, 39 participants were female and 71 were male. All of the students were enrolled in a required English class for listening and speaking skills for second year students. Students' scores on an English placement test given by the school determined which class they were placed into. With 100 points being a full mark, the scores ranged from moderately high for the advanced classes, to very low for the beginner classes. Five sections were opened for Engineering students, and four sections were opened for the International Culture students. The number of students in each class, as well as average mean scores on the placement test, is shown in Table 1.

Table 1

| | Dept.1 1 | Dept.1 2 | Dept.1 3 | Dept.1 4 | Dept.1 5 | Dept.2 1 | Dept.2 2 | Dept.2 3 | Dept.2 4 |
|---|---|---|---|---|---|---|---|---|---|
| Class size | 40 | 40 | 38 | 37 | 35 | 29 | 38 | 38 | 26 |
| test scores (Max. = 100 points) | 60 | 47 | 40 | 33 | 26 | 60 | 39 | 32 | 24 |

N=321  Dept. 1 = Engineering, Dept. 2 = International Culture

Due to time and place constraints, random assignment to groups was not possible. Therefore pre-testing was carried out to measure for statistically significant differences in cloze test performance among the groups. Based on the results of the pre-tests, four classes were selected to participate in the experiment as is shown in Figure 1. The differences in *n* size and actual class size were due to absences on the day of the test.

Figure 1

| | *Dictionary* | |
|---|---|---|
| Pair Work (*Discussion*) | + + (1)<br>In pairs w/dictionary<br>Dept 1, Class2<br>(n=28) | + - (2)<br>In pairs, no dictionary<br>Dept. 1, Class 4<br>(n=28) |
| | - + (3)<br>Alone w/dictionary<br>Dept .1, Class 3<br>(n=23) | - - (4)<br>Alone, no dictionary<br>Dept 2, Class 2<br>(n=31) |

N=110

For future reference the numbers in the upper left of each part of the table shall be used in describing and differentiating the groups, as follows:
(1) pairs-with dictionary, (2) pairs-no dictionary, (3) alone-with dictionary,
(4) alone no dictionary.

**Dictionary survey**

For this study the use of dictionaries was limited to electronic dictionaries. To determine how available electronic dictionaries were to students a survey was taken. It was found less than 40 percent of the students had bona fide electronic dictionaries, such as the Casio EX Word. However, all the students had access to cell phones that could use either dictionary applications, or could connect to online dictionaries. It was therefore decided to allow the use of both types of electronic devices.

## Pre-testing

In the initial stages of the study all nine classes participated in five pre-testing sessions, which used various cloze tests. The purpose of the pre-testing was threefold. First of all it was necessary to find out if four of the nine classes performed similarly enough to be considered equal. It was also useful in checking how participants might perform under all the various possible test settings, and in helping then feel familiar with cloze testing. Each time a pre-test was given, all the participants took it using the same format. For the first test, for example, all the participants worked alone with no dictionary. For the second, all the classes worked alone with a dictionary. In the end, all the classes experienced taking a test under all the levels of both independent variables. The use of pre-testing to help familiarize the participants with procedures and prepare them for taking the test seemed effective. Finally, pre-testing made it possible to work on adjusting the level of the test to avoid floor and ceiling effects. After each pre-test, the scores of all nine of the classes were analysized by performing either an ANOVA or Kruskal-Wallis, and then running post hoc tests to see if the means of the classes were statistically different. It became obvious that it would be difficult to find four classes that consistently performed at the same level, however, after completing five sessions of pre-testing, four classes were selected for the final phase of the study, based on the fact that the differences in their final pre-testingperformance was not found to be statistically significant. See Appendix 1. The final assignment of each class to one of the treatments was random.

## Materials

After selecting the groups to participate in the experiment, a single cloze test was administered. The topic was the history of popular music in the United States, and it was based on a factual text that had been written using 700 headwords. According to the publisher, headwords are defined as the words that form headings in a dictionary. Such texts, therefore, may include any of the derived forms in the associated word family. The choice to use a text written at the 700 headword level was made after examining the results of pre-tests prepared at the 1400, 1000, and 400 headword levels. The length of the passage was 260 words, and on average, every 7th or 8th word was deleted. In all, the paragraph contained 25 deletions. Although the optimum length of a cloze test ultimately depends on both the students and the test itself, a 25-deletion test is generally considered to be the most reliable in terms of length (Brown, 1988).

An example of the actual text appears in Appendix 2. The passage was printed and handed out to the participants, who completed it by filling in the deletions by hand. Twenty minutes were allowed to complete the test.

**Procedure**

All cloze tests were given during the first 20 minutes of each class. As each class had a different teacher, it was necessary to explain and standardize procedures. Short meetings were held before each testing session, with brief written instructions to explain specifics about that day's test. Upon completion of the tests, the teachers again met briefly to discuss any issues that may have come up, and to make observations or suggestions. Upon completing the pre-testing stage of the study, both the test administrators and the participants showed a good grasp of the testing procedure and seemed familiar with all aspects of the design. The assignment of participants to study in pairs was made randomly using a deck of cards.

In order to ease grading and bring a sense of closure to each test, peer grading was used. Participants were asked to exchange papers with someone they felt comfortable working with, and were then given a single answer sheet to share. There was always a lot of interest shown by the students, who actively discussed the answers and their results. They seemed to appreciate getting this kind of rapid feedback. Participant grading, however, was not always accurate and scoring had to be checked before entering the data into SPSS.

**Results**

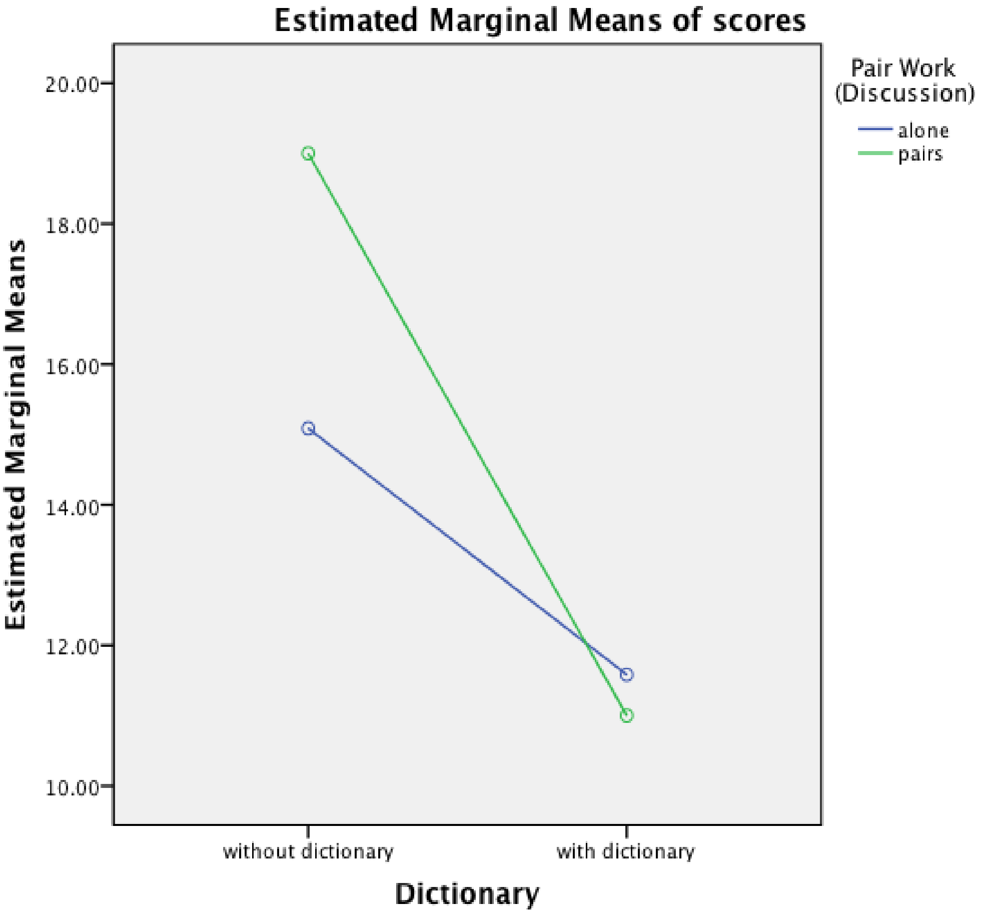
A 2x2 factorial ANOVA was used to calculate the interaction and main effects of dictionary use and Pair work (discussion) on cloze test performance. Table 2 shows that there was a significant interaction between the effects of dictionary use and discussion, $F$ (1,106) =7.74, $p$= .006, partial eta$^2$ = .068. The interaction is shown in the plot profile of the estimated marginal means of scores, Figure 2. Both the main effects of dictionary use, F(1,106) = 50.74, $p$ = .000, and discussion, F(1.106) = 4.256, $p$ = .04, were significant.

Table 2

Two-way Analysis of Variance for Cloze-test Performance as a Function of Dictionary Use and Discussion

| Source | *df* | *MS* | *F* | *p* | *Partial Eta Squared* |
|---|---|---|---|---|---|
| Dictionary | 1 | 899.64 | 50.74 | .000 | .32 |
| Discussion | 1 | 75.46 | 4.26 | .042 | .04 |
| Dictionary * Discussion | 1 | 137.22 | 7.74 | .006 | .07 |

Figure 2



The assumptions of independent observations, homogeneity of variances, and normal distributions of the dependent variable for each group were checked. The assumption of homogeneity of variances was not violated. The assumption of normal distributions of the dependent variable for each group was violated in the case of group (2) participants working in pairs without a dictionary, (Shapiro-Wilks) W = 0.01, p< .05. Figure 3 shows that the data for participants working in pairs without a dictionary (skewness-.117, kurtosis -1.441) is somewhat flat, with spikes in score clusters just above and just below the mean.  In Figure 4 the shape of the distribution was examined with a Q-Q plot which indicates that an assumption of normality was not unreasonable.  As the factorial ANOVA is not extremely sensitive to moderate deviations from normality (Glass et al. 1972,), it was decided to rely on the results of the factorial ANOVA for analysis of the interaction and main effects of the study.

Figure 3

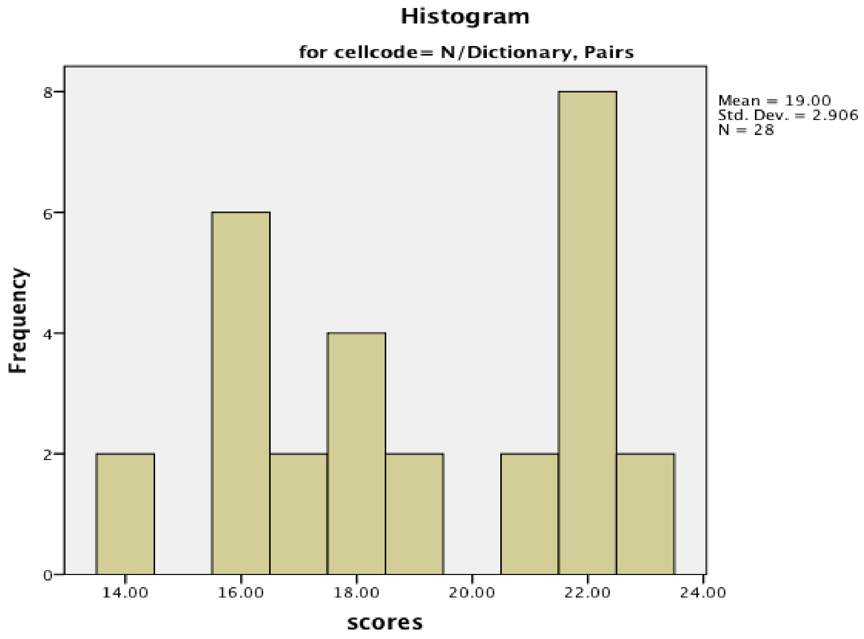
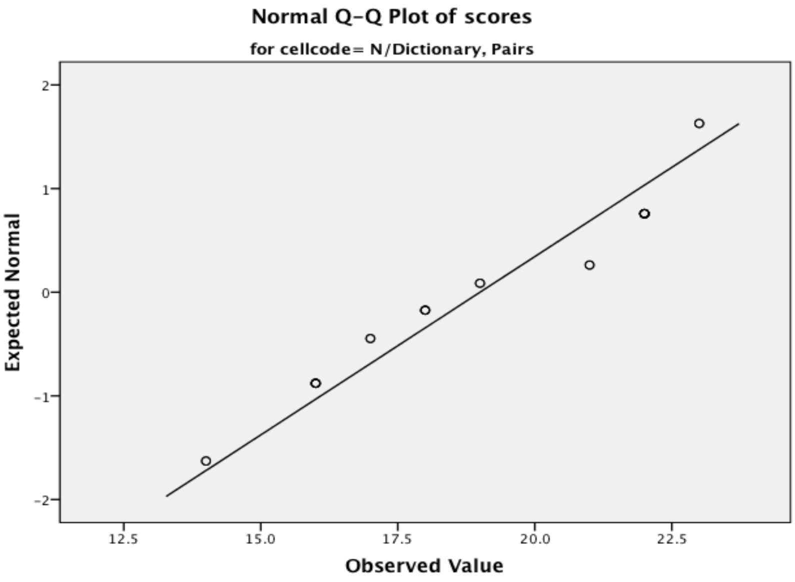


Figure 4



**Simple main effects**

A Tukey post hoc test was conducted to examine the simple main effects of the different levels of the independent variables. The post hoc multiple comparison gave the following results:

(1) Pairs- With Dictionary & (2) Pair – No Dictionary, *p* = .000, *d* = 2.38

(1) Pairs –With Dictionary & (3) Alone – With Dictionary, **NS** (not sig.)

(1) Pairs –With Dictionary & (4) Alone – No Dictionary, *p* = .004, *d* = 0.92

(2) Pair – No Dictionary & (3) Alone – With Dictionary, *p* = .000, *d* = 1.85

(2) Pair – No Dictionary & (4) Alone – No Dictionary, *p* = .007, *d* = 0.95

(3) Alone – With Dictionary & (4) Alone – No Dictionary, *p* = .016, *d* = 0.70

The cell size (n), the means, and standard deviations are presented in Figure 5. Other results of the post hoc multiple comparisons, including effect size, are contained in Table 3. Figure 6 shows the effect sizes of the simple main effects imposed on the profile plot of the estimated marginal means of scores.

Figure 5

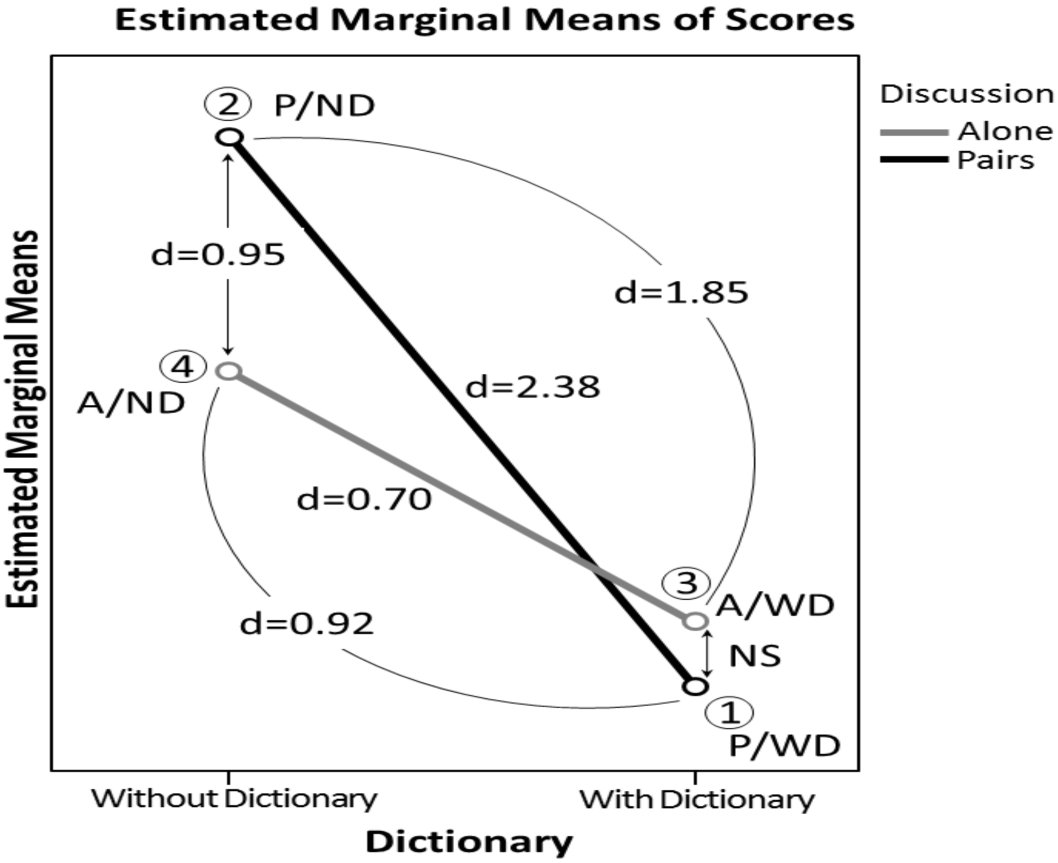
| | *Dictionary* | | |
|---|---|---|---|
| *Pair work (Discussion)* | + + (1)<br>In pairs w/dictionary<br>Mean = 11.00,<br>Std. Deviation = 3.76<br>(n=28) | + - (2)<br>In pairs, no dictionary<br>Mean = 19.00<br>Std. Deviation = 2.90<br>(n=28) | Row Mean = 15 |
| | - + (3)<br>Alone w/dictionary<br>Mean= 11.60<br>Std. Deviation = 4.86<br>(n=31) | - - (4)<br>Alone, no dictionary<br>Mean = 15.09<br>Std. Deviation = 5.05<br>(n=23) | Row Mean = 13.35 |
| | Column Mean = 11.3 | Column Mean = 17.05 | GRAND MEAN =14.17 |

Table 3

Multiple Comparisons
Tukey HSD
Dependent Variable: Score

| Cell Code | Cell Code | Mean Difference | Std. Error | Sig | *d* |
|---|---|---|---|---|---|
| 1) Pairs -W/Dic. | 2) Pairs- N/Dic. | -8.00 | 1.13 | .000 | 2.38 |
| | 3) Alone-W/Dic. | -0.58 | 1.10 | .952 | |
| | 4) Alone -N/Dic. | -4.09 | 1.18 | .004 | 0.99 |
| 2) Pairs- N/Dic | 3) Alone-W/Dic. | 7.41 | 1.10 | .000 | 1.84 |
| | 4) Alone -N/Dic. | 3.91 | 1.85 | .007 | 0.95 |
| 3) Alone-W/Dic. | 4) Alone -N/Dic. | -3.51 | 1.16 | .016 | 0.70 |

Figure 6



## Discussion

### Accepting or rejecting the null hypothesis

The interaction of the two main effects of the study made evaluating the assumptions difficult, and it was sometimes necessary to qualify the decisions made.

In the case of the first hypothesis, $H_o^1$: There will be no significant interaction between the main effects of dictionary use and discussion, we fail to reject the null hypotheses.

In the case of the second hypothesis, $H_o^2$: There will be no significant difference between the test scores of participants using dictionaries and those not using dictionaries on the cloze test, we reject the null. Test scores of participants not using dictionaries were always higher.

In the case of the third hypothesis, $H_0^3$: There will be no significant difference between the test scores of participants working in pairs and participants working alone on the cloze test, the decision of whether or not to reject the null hypothesis is confounded by the significant interaction between the main effects of dictionary use and discussion. The decision to accept or reject the null hypothesis had to be made case by case based on the multiple comparisons of the simple main effects The results for (1) pairs-with dictionary and (3) alone-with dictionary were not statistically significant, and we fail to reject the null. All other combinations of the two levels of independent variables were statistically significant, and all showed a very large effect size. In particular, in the cases of (2) pairs-no dictionary and (3) alone-with dictionary, and (2) pairs-no dictionary and (4) alone-no dictionary, it was possible to reject the null, as pairs always outscored participants working alone. Finally, (4) alone-no dictionary scored higher than (1) pairs-with dictionary, however, and the difference was statistically significant.

### The meaning of the interaction

The results showed that there was an interaction between the different levels of the independent variables. The nature of an interaction is “that the effect of one factor depends on conditions controlled by the other factor", (Yatani, 2010). To better define the interaction, a profile plot of dictionary use and pair work (discussion), shown in Figure 2, was produced. This plot shows that (2) pairs-no dictionary scored higher on cloze tests than (4) alone-no dictionary. However, when both pairs and individuals used a dictionary, then (3) alone-with dictionary will score higher than (4) pairs-with dictionary.

One way to explain the interaction would be to say that when working with a dictionary, the mean score of individuals was higher than that of pairs, but not enough to be statistically significant. However, when working without a dictionary, the mean score of pairs was higher than that of individuals, and this difference was statistically significant. The negative effect of dictionary use is especially strong when participants are working in pairs. Because of the interaction, it cannot be said that working in pairs (discussion) always led to higher test scores than working alone.

**Examination of simple main effects**

Unimpeded discussion, condition (2) pairs-no dictionary, is more effective in raising test scores than any other possible combination of the levels of the independent variables. In other words, nothing is more effective at generating higher tests scores than working in pairs without a dictionary. This would seem to be the clearest and most outstanding finding of the study. But why is this so? To answer that question is it necessary to examine all of the simple main effects and engage in some grounded speculation.

The only simple main effect that was not statistically significant was (1) pairs-with dictionaries and (3) alone-with dictionaries. The fact that these two groups had the lowest scores, and were significantly different from (2) pairs-no dictionary and (4) alone-no dictionary implies that the act of dictionary accommodation has an affect on levels of performance on cloze tests. There are several reasons why this may be the case. One of the most important is time. The test was limited to twenty minutes. Time spent working with a dictionary could actually be thought of as time lost. In other words, by working alone with no dictionary, or by talking about the test problems with no dictionary, participants could spend more of the allotted time solving the test problems than the participants who used dictionaries.

The results also cause one to think about the role of the dictionary. Since EFL reading tests involve comprehension, a dictionary, which is intended to clarify word meanings, translations, and descriptive grammar points, may not be of much help. Attempts to clarify context, such as discussion, on the other hand would help provide test takers with more useful clues (Bennsoussan, 1983). Another dictionary related issue reason may be differences in participants' skill in using dictionaries. Levels of dictionary skills were not measured for this test, but it would be reasonable to imagine that not all students share the same level of dictionary familiarity and skill, and that the test performance of students with lower levels of dictionary familiarity and skill would be further hampered under the condition of dictionary accommodation.

Two other important assumptions may be put forth regarding the statistically significant higher test scores of (1) pairs–with dictionary and (2) alone–no dictionary. Both of these assumptions are related to the role constructivist learning theory plays in pair work, dictionary use, and cloze testing. The first assumption involves the participants' application of knowledge towards the comprehension of test items. In accordance with Bloom's Taxonomy of Learning, the participants in both of the higher scoring groups, seem to be moving beyond a narrowly based focus on factual knowledge. It would seem that they are making use of conceptual knowledge to better understand and come up with better answers to the test problems (Krathwohl, 2002). Thus, in the case of participants working alone without a dictionary, it would seem that they are benefiting not only from spending more of the allotted time working on the test, but also from applying what they know towards comprehending and solving the test problems. As opposed to spending time getting more information or knowledge from an outside source, they are using higher order thinking skills to apply what they already know to raise their level of understanding. This would seem to demonstrate the greater effectiveness of using higher order thinking skills on tests. But what are we to make of the even higher, statistically significant performance of (2) pairs-no dictionary? While these pairs are able to enjoy the same benefits of dictionary freedom as (4) alone-no dictionary, they are also able to take advantage of social constructivist learning. Since each participant will know something the other does not, it will put them in a situation where, by using discussion to engage in reciprocal teaching, scaffolding and collaborative leaning, they can assist each other in moving from not knowing to knowing (Sinky, 2010). They have the benefit of engaging with the test in a way that allows for an exchange of information and increases their knowledge by sharing. Moreover, they seem to be coming up with new and unique ways to apply what they know and comprehend, and then going on to solve more test items in a way that participants working alone cannot.

## Limitations of the study

There were several limitations to the study. The most important had to do with the scheduling of the final cloze test. Unintentionally, the day the final cloze test was given coincided with a minor national holiday that was not observed by the university. Although school was officially in secession, and classes were held as usual, there were a large number of absences. Of the original 144 participants scheduled to take the test, 34 participants were absent, representing a 24% absentee rate. This resulted in a decrease in sample size. By the time the conflict was discovered there was no opportunity to reschedule or re-administer the test, so the results had to be calculated with the smaller sample size of 110. Furthermore, the effect of the 24% absentee rate on the pre-test based statistical analysis to determine the non-significance in levels of ability is not clear.

Although efforts were made to ensure that all the classroom teachers were administering the tests in a similar fashion, it is likely that there were differences. These differences may have influenced participant performance. While the aim was to have a well informed, competent, but neutral administration of the test, it is questionable if this happened equally in all cases.

Finally, every effort was made to ensure that the level of difficulty of the final cloze test was similar to that of the final pre-test that was used to choose the participants. However, even though they dealt with similar content and used the same headword count, the two tests may have had slightly different levels of difficulty. Unfortunately, to validating the tests to make sure that they were of equal difficulty would have demanded more time than the study allowed.

## Implications of the study

First, one must allow for the negative effect of dictionary use in lowering all participants' test performance. Then it can be said that the results indicate working in pairs, and engaging in discussion is a better way to allow students to perform at higher levels than having participants work alone. The implication of this is important for classrooms. Since participants are learning while working together, and creating an outcome that they could not reproduce working alone, pair work and discussion can be said to be have an important role to play in making lesson plans, carrying out projects, and working on assignments.

It is possible to assess pair work fairly, however, not in the same way one would assess individual work. For this test, the work of two participants was used to make one, shared score. But in fact, what is being assessed is the participants' ability to use cooperative learning and social constructivist constructs to produce a unique result, which is different from assessing their initial individual level of skills, or their ability to preform on tests working alone.

This study did not attempt to predicate individual skills based on paired assessment. It was also not possible to know about differences in skill among participants working in pairs. Ultimately, how a pair performs depends on who is placed in the pair. Change one of the partners, change the result. In this way, assessing the results of pair work is more complicated, and perhaps a bit more elusive than assessing individual performance.

It would seem that pair assessment might augment, but not replace individual assessment, and that the real value of pair work lies in its potential for better learning and higher levels of performance.

Another implication of the study is the need for more explicit vocabulary instruction for the students involved in the study, students who are doubles representative of many, but certainly not all Japanese university students. In order to make the test viable, it was necessary to go through a two-week process of constantly reducing the level of difficulty. The first pre-tests used authentic English materials that proved to be much too difficult for the participants. The move was then made to slightly advanced graded readers. The downward spiral in the level of English continued as tests were made using headword counts of 1,400, then 1000, and eventually 700 words. This is a very low threshold for second year university students and does not bode well for students' chances of success in using English outside of school, in the real world. The study showed the ineffectiveness of using dictionaries on cloze tests to rose test scores, but that does not mean that dictionaries would not be useful in instruction and long term, explicit vocabulary learning, an area that the students in this study would benefit from.

Levels of dictionary familiarity, dictionary use strategies, and dictionary skill were not addressed in the study, but they are expected to have some bearing on students' overall vocabulary ability. It would be interesting to see if, after instruction and practice in using dictionaries, vocabulary quizzes could be designed and administered that would better reflect a participants command of dictionary skill.

## References

Albus, D., Thurlow, M., Liu, K., Bielinski, J. (2005). Reading test performance of English-Language-Learners using a dictionary. *The Journal of Educational Research, Vol. 98, No. 4*. 245-254.

Aust, M. Kelly, M.J., Roby, W. (1993). The use of hyper-reference and conventional dictionaries. *Educational Technology and Development, Vol. 41. No. 4,* 63-73.

Bensoussan, M., Sim, D. Weiss, R. (1984)The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Modern Language Journal Vol. 68, Is. 3,* 230-239.

Bensoussan, M. (1983) Dictionaries and tests in EFL reading comprehension. *ELT Journal* v 37, n 4, 341-345.

Brown, J.D. (1988). Cloze tests and optimum test length. Shigeken: JALT Testing & Evaluation SIG Newsletter, 2 (1) October 1998, 18-21.

Brown, J.D. (2002). Do cloze tests work? Or is it just an illusion? *Second Language Studies, 21*(1), Fall 2002, 79-125.

Chapman, M. (2007). The effectiveness of cloze testing for Japanese Learners of English. *JALT Hokkaido Journal* Vol. 7. 24-34.

Glass, G.V., P.D. Peckham, and J.R. Sanders. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. Rev. Educ. Res. 42: 237-288.

Hadley, G. S. NaaykensJ.E. (1997). An Investigation of the selective deletion cloze test as a valid measure of grammar-based proficiency in second language learning. *Niigata University Linguistics and Culture research, 3,*December 1998, 111-118

Johnson, D. W., Johnson, R. T. (2009) An educational psychology success story: social independence theory and cooperative learning. *Educational Researcher, vol 38, No, 5.* 365-379.
DOI 10.3102/0013189X09339057

Knight, S. (1994). Dictionary use while reading,: the effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal* Vol 78. No. 3, 285-299.

Saeedi, M. Tavakoli, Rashimi, S. Kazerooni, Pavaresh, N. (2001). Do c-test and cloze procedure measure what the purport to be measuring? A case of criterion-related validity. *World academy of Science and Technology* 50, 1034-1043.

Scholfield, P. (1982). Using the dictionary for comprehension. *TESOL Quarterly* Vol 16, No. 7, 185-194.-

Storch, N. Aldosari, A. (2013). Pairing learners in pair work activity. *LANGUAGE TEACHING RESEARCH,* 2013, 17. 31-48,

Yatani. J. (2011). Statistics for HCI Research. In *yatani.jp*. Retrieved October 20, 2013, from http://yatani.jp/HCIstats/HomePage.

Sinky, Z. (2010) Social constructivism and cognitive development theory. In *Slidedhare.net,* Retrieved October 28, 2013, from http://www.slideshare.net/sinkyzheng/social-constructivism-cognitive-development-theory.

(Submitted: February 10, 2014; Accepted: March 24, 2014)

Appendix 1

Results of final pre-test

**Multiple Comparisons**

Tukey HSD[a,b,c]

| class | N | Subset 1 | Subset 2* |
|---|---|---|---|
| D2-3 | **9** | **5.44** | |
| D1-4 | **11** | **7.55** | **7.55** |
| D2-2 | **9** | **8.22** | **8.22** |
| D1-3 | **16** | **8.69** | **8.69** |
| D1-2 | **10** | | **11.40** |
| Sig. | | **.152** | **.058** |

*Subset 2 was used for the final test. D1-2, D1-3, D1-4 and D2-2 were randomly assigned as follows:

Dept 1-4= Group 1) Pairs, With Dictionary

Dept 1-2=Group 2) Pairs, No Dictionary

Dept 1-3=Group 3) Alone, With Dictionary

Dept 2-2= Group 4) Alone, No Dictionary

Appendix 2

Final Cloze Test - vocabulary taken from a list of 700 headwords.

Name ___________________________ Number ____________________________

Name ___________________________ Number ____________________________

SCORE___________________

Word List
kinds rhythm young they made a where had the bands States music to story fathers looking began did he changed America was different listen in

Young people all over the world love pop music and listen to it almost everyday. But how did pop music start? 1) ______ did it come from? This is the 2) ____________ of how pop music began in the United 3)_______________. Sixty years ago there were many different 4) _________ of popular music: black music, white music, gospel, 5)________ of the church. It was a time of world peace and 6)________ was rich. Young people had jobs. 7) ____________had money to buy radios and to 8)________________ to records. They could buy electric guitars. They 9)________________ to play the new music called 'R and B', or 10) ________________and blues, and also rock and roll. Many 11) _______ ______ and singers became popular and famous. But 12) ________ most famous of all was Elvis Presley. Elvis 13) ____________ listened to different kinds of music when14) _____ ____________ was a boy. He was born in 1935. He came from 15)________ poor white family. He often heard black people playing 'R and B'. He 16) _____________ his first record in 1954 and 17)_________ 1956 he had his first number one hit. Suddenly he was famous and rich. He was good 18)________________, he was a great singer, and he19) _ ____________ different. Girls liked him. Boys wanted 20) _______ be like him.

Before Elvis, most 21) __________ people were like their mothers and 22) ____ _______ . They wore the same clothes, they 23) __________ the same things. After Elvis, everything 24) ___________________. Young people wanted to be different. They wanted

25) __________ clothes. They had different music. Elvis was the first star of rock and roll, the first pop star.