

JALT Journal

JALT Journal is the research journal of the Japan Association for Language Teaching (JALT). It is published semiannually, in May and November. As a nonprofit organization dedicated to promoting excellence in language learning, teaching, and research, JALT has a rich tradition of publishing relevant material in its many publications.



Links

- JALT Publications: <http://jalt-publications.org>
- *JALT Journal*: <http://jalt-publications.org/jj>
- *The Language Teacher*: <http://jalt-publications.org/tlt>
- *Conference Proceedings*: <http://jalt-publications.org/proceedings>

- JALT National: <http://jalt.org>
- Membership: <http://jalt.org/main/membership>

Provided for non-commercial research and education.
Not for reproduction, distribution, or commercial use.

Articles

Effects of Preparation and Use of Keyword Lists on a Classroom Story-Retelling Test

Hidetoshi Saito
Ibaraki University

The purpose of this quasi-experimental study was to explore the effects of test practice and keyword use on story-retelling test performance under four conditions. Eighty-two beginning and intermediate Japanese university students enrolled in English courses were required to practice orally summarizing two passages using a keyword list and also instructed to orally summarize one of two previously unseen passages without preparation. In the test session, two groups experienced two conditions that were identical and one that was different. Both groups retold one practiced passage with keywords at hand and a new passage without a keyword list. Group 1 retold another practiced passage with the keyword list withheld, whereas Group 2 read an additional new passage, made a keyword list, and retold it with the keyword list but without practice. Test practice was found to improve performance, but keyword list use induced better performance only when used with practice.

テスト準備とキーワードリストは口頭要約テストに役立つか。この研究はテスト準備とキーワードリストの使用が口頭での要約テストに役立つかを調査することを目的とした。日本人大学生（初中級者）2グループ（計82名）が二つの同一条件と一つの異なる条件でそれぞれ英文要約を行った（計三条件づつ）。参加者は予め二種の英文が渡され、キーワードリストを作って練習をするように指示された。また、その場で新しい英文の要約を行うことも指示された。試験当日、両グループともまず練習した英文をリストとともに要約し、その後新しい英文の要約も行った。グループ1はさらに、準備したキーワードリストなしで練習した英文の要約を行った。グループ2はその場で新しい英文を読み、キーワードリストを作って

要約を行った。結果として、練習したほうが、練習をしていない場合より良いが、キーワードリストは練習した場合のみに有効であることがわかった。

Opportunity for practice is critical in learning a new language. Repeated practice is required to automatize skills (Anderson, 1999). Despite the significant role of practice in skill learning, practicing for tests has been considered inappropriate because of purported inflation of test scores without actual improvement in the target skills or knowledge (Lai & Waltman, 2008). From a classroom learning perspective, however, preparing for performance tests, such as speaking and writing tests, is widely believed to have a positive influence on student engagement in learning in and outside the classroom, especially in the context of foreign language learning, where learners have little opportunity to use the target language in daily life.

To explore whether test practice and preparation improve classroom test performance, this study focused on the effect of practice and use of a self-prepared keyword list on oral test performance. Two hypotheses relevant to the role of practice and keyword use in the classroom-testing context were proposed: the test practice effect (TPE) hypothesis and the keyword use effect (KUE) hypothesis.

Literature Review

The first hypothesis in this study is the TPE hypothesis, which states that test practice of the target performance (rather than no practice) will facilitate performance on speaking test tasks. Previous studies in second language acquisition (SLA) have consistently shown that repeated practice improves the learner's speaking performance on a task (Arevart & Nation, 1991; Bygate, 2001; Bygate & Samuda, 2005; Gass, Mackey, Alvarez-Torres, & Fernández-García, 1999; Kawauchi, 2005; Lynch & Maclean, 2000; Williams, 1992). The TPE hypothesis is consistent with these studies, revealing that practice or repeated implementations of the target activity (defined by DeKeyser, 2007) improve current skills and hence facilitate language performance.

One explanation for the practice effect is that previous experience with the same task type (familiarity of task structure) and content (prior knowledge; Mackey, Kanganas, & Oliver, 2007; Skehan, 1998) results in reduced cognitive load and thus efficient processing of the conceptualization and formulation of speech (Bygate & Samuda, 2005). Better performance in combinations of accuracy, fluency, and complexity can reflect efficiency in cognitive processing (Robinson, 2005; Skehan, 1998). Efficient cognitive

processing is made possible by bypassing access to declarative knowledge and combining those declarative (or descriptive, explicit) rules to form a new, simpler rule (Anderson, 2007).

Although SLA research supports the practice effect, testing studies have shown little support for the effects of preparation or coaching for large-scale American and British academic tests, including the Cambridge FCE, GRE, IELTS, SAT, TOEFL, and other academic aptitude tests (Alderman & Powers, 1980; Bachman, Davidson, Ryan, & Choi, 1995; Green, 2007; Nguyen, 2007; Powers, 1985, 1986, 1993). However, because the preparation programs compared and aggregated in each study varied in length, teaching methods, and content, nil effects based on noneffective preparation in one program were likely to offset potential benefits of effective preparation in another program. A recent, well-controlled large-scale EFL test preparation study (Xie, 2013) showed a small but statistically significant effect of drilling on test scores. This implies that the efficiency of test preparation depends on the extent to which preparation matches the actual test task—whether what is rote-learned is on the test. In fact, the degree of similarity between an actual assessment and classroom instruction may account for this apparent contradiction among research studies (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). The effect of test preparation on large-scale, norm-referenced, academic tests could be smaller because what was tested may have been “distal/remote” from what had been instructed in lessons and hence insensitive to student learning. On the other hand, the effect of practice in SLA studies is larger because what the participants repeatedly practiced is identical to what was tested in the studies. In cognitive psychology, this is known as transfer-appropriate processing: learners best transfer skills learned in a certain context to a similar context (Lyster & Sato, 2013).

Although repeated practice potentially boosts immediate test performance, the use of a keyword list—commonly employed in public speaking—may also improve oral performance. However, none of the previous studies in these areas has tested the effects of the availability of a keyword list. Authors of several textbooks on teaching formal speechmaking (e.g., Gregory, 2002; Nation & Newton, 2008) have asserted that brief speaking notes not only encourage natural speaking but also provide the speaker with cues and a sense of security. The keyword use effect (KUE) hypothesis proposed herein states, based on a logical extension of these claims, that the use of a keyword list improves oral test performance. Joe (1998) compared the effect of the presence of the original passage, rather than keywords, on a story-retelling task. In Joe’s study, after a 15-minute practice, two groups

orally summarized a passage, but only one had access to the passage while summarizing. Although both outperformed a nonpracticing control group in vocabulary recall, there was no difference between the two practice groups. This suggests that practice, not access to the passage, is important; in other words, once practiced, vocabulary recall no longer benefits from the availability of the passage while retelling. If generalized to the present context, this is contrary to the KUE hypothesis and the belief regarding the usefulness of keywords in story retelling and speech.

Questions still remain regarding the extent to which practiced performance differs from impromptu performance and whether keyword use boosts performance. Furthermore, it is of great interest to examine whether the effects of practice and keyword use are additive.

Hypotheses of the Present Study

Four different conditions to examine the plausibility of the two hypotheses were set up (see Table 1). In the fully assisted condition, participants had a chance to practice and use a keyword list for the task. In the practiced condition, participants had a chance to practice and make a keyword list but did not have access to it for the task. In the keyword-assisted condition, participants did not have a chance to practice, but they did make a keyword list and had access to it during the task. Finally, in the impromptu condition, participants did not have a chance to practice nor to make a keyword list for the task. Performance under these four conditions was compared and predictions generated from the TPE and the KUE hypotheses, as described in Table 2, were tested.

Table 1. Conditions in the Present Study

Conditions	Availability	
	Practiced	Keyword list
1 Fully assisted	✓	✓
2 Practiced	✓	0
3 Keyword-assisted	0	✓
4 Impromptu	0	0

Table 2. Predictions of the TPE and KUE Hypothesis

Contrast	Condition	Hypothesis		Condition
		Test practice effect	Keyword use effect	
1	Fully assisted	=	>	Practiced
2	Fully assisted	>	=	Keyword-assisted
3	Fully assisted	>	>	Impromptu
4	Practiced	>	<	Keyword-assisted
5	Practiced	>	-	Impromptu
6	Keyword-assisted	-	>	Impromptu

Note. > indicates that the condition in the left column is predicted to induce better performance compared to the condition in the right column; = indicates that performances on both conditions are predicted to be equivalent; - indicates that the hypothesis cannot predict performance.

As shown in Table 2, the KUE hypothesis predicts the advantages of keyword use in Contrasts 1, 3, 4, and 6. The KUE hypothesis lacks a prediction regarding Contrast 5 (practiced vs. impromptu) because it does not concern the keyword list, and it predicts equal effects in Contrast 2 (fully assisted vs. keyword-assisted) because of the availability of a keyword list in both conditions.

The TPE hypothesis predicts the advantages of test practice over nonpractice in Contrasts 2, 3, 4, and 5. However, because the TPE hypothesis does not consider the role of a keyword list, predictions for Contrasts 1 and 6 differ from those of other contrasts. The TPE hypothesis lacks a prediction regarding Contrast 6 (keyword-assisted vs. impromptu) because it does not concern practice. Concerning Contrast 1 (fully assisted vs. practiced), it can be assumed that repeated practice may reduce the need for keyword list assistance, as in Joe's (1998) study on oral test performance. A strong version of the TPE hypothesis could then mean that practicing facilitates better performance by participants on the test task, and when ample practice is implemented, a keyword list does not further improve test performance.

Method

Design

This is a repeated-measures study. All participants took a standardized speaking test (Telephone Standard Speaking Test [TSST]; ALC Press, 2016), after which they were exposed to three task conditions—fully assisted, impromptu, and either practiced or keyword-assisted—of a story-retelling task using different passages. Performance in the impromptu condition served as a baseline to which the performance scores of the three other conditions were compared. Critical comparisons were made within subjects; baseline and target conditions were compared *within* each group, rather than to control or experimental groups. However, one between-group comparison (the practiced condition vs. keyword-assisted condition) was necessary because participants in each group performed only one or the other of these conditions.

Participants

Eighty-two native-speaking Japanese university freshmen and sophomores (34 males and 48 females, aged 18-21) participated in the study. Group 1 ($n = 29$), was made up of students majoring in education who were enrolled at a national university in an English course that met for 15 weeks once a week for 90 minutes. Group 2 ($n = 53$) consisted of students in two classes, one majoring in education ($n = 20$) in a different section of the same course as Group 1 and the other majoring in humanities and science ($n = 33$) at the same university in a general English course that met for 15 weeks twice a week for 90 minutes. The main purpose of these courses was to improve discussion and debate skills. All participants took the TSST approximately 2 weeks before the story-retelling test. Table 3 shows that most participants in both groups performed at Levels 3 and 4 of nine possible levels, except for six students in Group 2 who performed at Levels 5 and 6. The TSST score was regarded as a proficiency factor in the present study and entered as a moderating variable, as in previous studies (e.g., Kawauchi, 2005).

Table 3. The Speaking Proficiency Levels of Participants

Proficiency level	Group 1	Group 2
3 (Novice high)	9	28
4 (Intermediate low)	20	19
5 (Intermediate low plus)	0	5
6 (Intermediate mid)	0	1
Total	29	53

Materials

The three reading passages for this task were selected from a university-level EFL textbook (Day & Yamanaka, 1998). The passages covered interracial marriage, gay rights, and notification to cancer patients, all of which were appropriate debate topics for the participants, whose mean English proficiency was low intermediate or high beginning. Four readability indices confirmed an approximate equivalence of surface linguistic readability of the three passages (see Appendix B). Two graduate students and two university professors of English examined the comprehensibility of the passages and agreed unanimously that all passages were approximately equally comprehensible.

Procedure

Given the course objectives, the use of a story-retelling test—oral summary of a reading passage—was considered appropriate, and all participants practiced the task in pairs four times with a self-prepared keyword list guided by comprehension questions on the worksheet. Earlier studies have also suggested that story-retelling tasks facilitate the acquisition of L2 vocabulary (Joe, 1998) and grammar (Muranoi, 2007), possibly because the text provides repeated opportunities for the learner to encounter and use the target forms. Story retelling has also been a commonly used L2 test method (Brown & Abeywickrama, 2010; Underhill, 1987) and has been supported for its appropriateness for university courses (Hirai & Koizumi, 2009). The target construct here is skill in orally summarizing a short reading passage—a necessary preparatory skill for debate.

On the test, the participants were randomly assigned to one of the six possible sequences of the three reading passages (i.e., ABC, ACB, BAC, BCA, CAB, CBA). Assignment of passages was counterbalanced across participants.

One week before the test, the participants were given two new reading passages and were encouraged (on the grounds that the passages would be part of their final exam) to practice using a self-prepared keyword list. They were also informed that in addition to one of the two passages they practiced, they would also retell new passages, which would be chosen by the tester during the test.

During the test session, participants came to the instructor's room individually. As shown in Figure 1, with the exception of the second retelling, both groups underwent the same test procedures. In the first retelling task (the fully assisted condition), both groups orally summarized one practiced passage—chosen by the instructor during the test—using a keyword list. In the second retelling task, Group 1 retold the second passage, which they had also prepared for, but the prepared keyword list was withheld (the practiced condition). The participants were assured, to alleviate any surprise (because they had been told they would have to retell only one of the two practiced passages), that this particular performance would not be included in their final grade. In the second task for Group 2, they were given a new passage and asked to read it and make a keyword list within 5 minutes (the keyword-assisted condition). In addition, they were allowed to use the dictionary and spend an additional 2 minutes to prepare, if needed. In fact, all participants spent 7 minutes reading and making a keyword list. Participants then performed a story retelling using the newly made keyword list. In the third retelling task, both groups retold a new third passage given during the test, with no keywords available. They were given 5 minutes to read the third passage and were permitted to use a dictionary. All participants finished reading within 5 minutes, and 22 participants used the dictionary once or twice. Although there was no time limit for the retellings, nearly all participants completed each retelling within 5 minutes, the only exception being one participant who spent about 5 minutes and 30 seconds on two performances.

In this study, the order of conditions that each participant went through was fixed, because it was believed that participants could feel more relaxed if they started with what they had practiced before performing the impromptu condition. Counterbalancing of the order of conditions was avoided to elicit the best possible performance—that is, “biased for best” (Brown & Abeywickrama, 2010, p. 44). The fixed order of three conditions could possibly generate incidental practice and fatigue effects. However, an incidental practice effect—not the deliberate practice effect on which the present study was focused—was unlikely to occur, because all participants

had ample opportunity to practice the same task type before the test. Although the possibility of a fatigue effect could not be eliminated, the fact that the entire test session took only 25 minutes minimized this concern.

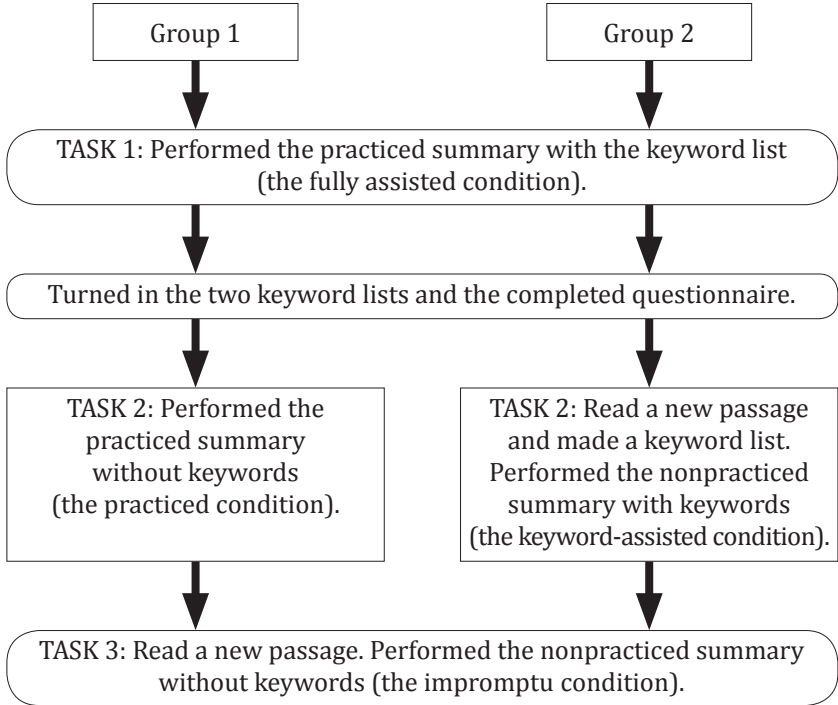


Figure 1. Test procedure of the present study. The only difference between Groups 1 and 2 in the test procedure was the second task.

Participants also completed an open-ended questionnaire investigating how they practiced for the test as well as how long they spent practicing. They returned the completed questionnaire along with the keyword lists to the researcher after the first story-retelling task. When necessary, the researcher asked questions after the test session to clarify participants' responses to the questionnaire. The researcher then asked each participant if he or she had ever read the new passage, to detect leakage from classmates, but all denied having done so.

All performances were audio recorded, and the randomized audio files were subjected to blind evaluation by three raters, all of whom had exten-

sive experience in teaching English at the college level in Japan and were trained for about 2 hours by the author before they started rating. They rated performances on three categories—language (mainly grammatical accuracy), fluency, and content (the adequacy of information covered)—using a 4-point scale (see Appendix A). This test was developed for the course and piloted twice before this research. The Rasch reliabilities of the second pilot test were .89 (examinees), .88 (raters), and .97 (items).

In this study, the participants themselves determined the extent to which they practiced. The instructor advised them to write keywords on the worksheet and to practice speaking aloud using the keyword list, as practiced in the lessons. The practice time thus was at each learner's discretion. It must be acknowledged that this lack of standardization in practice time weakened the internal validity of the study; however, it strengthened the study's ecological validity in that such individual variation in practice time reflected classroom reality.

Analysis

Because key statistical assumptions (normality, homogeneity of variance) were met, one between-subject and one within-subject (proficiency [TSST test scores] x condition [fully assisted, impromptu, and keyword-assisted or practiced]) analysis of variance (ANOVA) with Rasch performance measures as the dependent variable was run for each group to test the predictions. The Rasch analysis with the Facets program (Linacre, 2008) calibrates rater severity, item difficulty, condition difficulty, and participant performance measures. By taking into account such measurement factors (facets) as raters, items, and conditions, the Rasch analysis generates more plausible performance scores, which would not be possible with raw scores alone. The Rasch analysis allows for comparing measures across test facets on the common logit (log odds ratio) scale. Deviant raters, items, and participants are flagged through associated fit statistics for quality control. The Rasch model requires several strong assumptions, one of which is unidimensionality. The unidimensionality requirement means that all items on the same test measure the same dimension; misfitting items, persons, or both will return extreme fit statistics. The sample size of the present study, 82, was short of the recommended sample size of 120 (30 observations per factor) for achieving stability across samples (Linacre, n.d.), thus limiting generalizability. In addition, as a follow-up analysis, the potential mediating roles of the keyword list and the length of practice time for performance gain were examined through linear regression analyses.

Results

The Rasch analysis indicated high reliabilities of all four facets of raters, items, conditions, and participants (see Appendix C). The separation reliability value, 3.42, of participants indicates that participants can be divided into three groups. Table 4 displays descriptive statistics for the Rasch performance measures for the two groups. In both groups, two notable points are that, first, the differences between the fully assisted and impromptu conditions were greater than three logits with large effect sizes of $d = 1.79$ (Group 1) and 1.33 (Group 2). Second, the impromptu condition had the largest standard deviation, indicating a wide variation in performance. The Rasch analysis indicated a good fit of the data, with no misfitting items or raters. There was only one strongly misfitting participant based on the proposed criteria (Fisher, 2007), but this participant was retained for the subsequent analyses, because the purpose was descriptive for hypothesis testing rather than prescriptive for measurement construction. The performance measures generated from this Rasch run were then subjected to one between-subject and one within-subject (proficiency levels x condition) ANOVA.

Table 4. Means and Standard Deviations of Participant Performance Measures by Condition

Condition	Group 1 ($n = 29$)		Group 2 ($n = 53$)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Fully assisted	2.10	1.36	2.74	2.18
Practiced	0.48	1.74	-	-
Keyword-assisted	-	-	-0.29	1.97
Impromptu	-1.07	2.10	-0.46	2.58

Note. These are Rasch logit measures derived from raw scores.

Table 5 shows the results of the mixed-design ANOVA for Groups 1 and 2. The findings indicated statistically significant main effects of conditions and proficiency but no interaction effect for either group (Figures 2 & 3). This means that performances across conditions differed regardless of proficiency. The post hoc multiple comparisons for Group 1 indicated statistically significant differences among the three conditions (the fully assisted, practiced, and impromptu conditions). The follow-up analysis for Group 2

showed statistically significant differences between the fully assisted and the keyword-assisted conditions, but not between the keyword-assisted and the impromptu conditions. Because of the research design, these ANOVAs did not include a comparison between the practiced and the keyword-assisted conditions. A between-group *t*-test comparison of the two conditions was thus performed with the Bonferroni-adjusted probability set at .025. To remedy unequal sample sizes between the two conditions, 29 cases of those in the keyword-assisted condition were randomly sampled to match the sample size of the practiced condition. The result showed no statistically significant difference between the practiced and the keyword-assisted conditions, $t(56) = 1.57, p > .10$, effect size $r = .16$. To summarize the results, Table 6 displays all contrasts again. The TPE hypothesis is supported in Contrasts 2, 3, and 5; the KUE hypothesis is supported in Contrasts 1 and 3. Both are rejected in Contrasts 4 and 6.

Table 5. Results of Repeated-Measure ANOVAs for Groups 1 and 2

Source	Group 1				Group 2			
	<i>df</i>	<i>F</i>	<i>p</i>	η^2	<i>df</i>	<i>F</i>	<i>p</i>	η^2
Within subjects								
Condition (C)	2	24.44	.00	.47	2	14.92	.00	.23
Interaction (C x P)	2	0.34	.71	.01	6	0.79	.58	.04
Error	54				98			
Between subjects								
Intercept	1				1			
Proficiency (P)	1	4.40	.04	.14	3	6.72	.00	.29
Error	27				49			

Table 6. Predictions Supported by the Results of the Present Study

Condition	Hypothesis		Condition
	Test practice effect	Keyword use effect	
1. Fully assisted	=	> ²	Practiced
2. Fully assisted	> ¹	=	Keyword-assisted
3. Fully assisted	> ^{1,2}	> ^{1,2}	Impromptu
4. Practiced	>	<	Keyword-assisted
5. Practiced	> ¹	-	Impromptu
6. Keyword-assisted	-	>	Impromptu

Note. Superscript 1 indicates the results of Group 1 supports the prediction; superscript 2 indicates the same for Group 2.

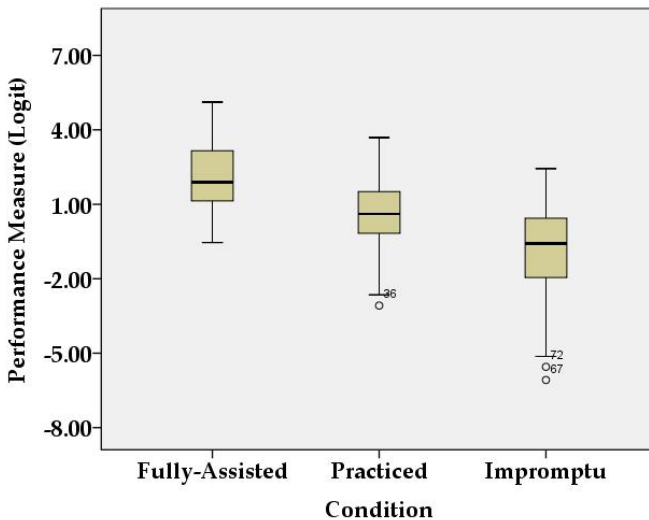


Figure 2. The boxplot of performance measures of three conditions for Group 1.

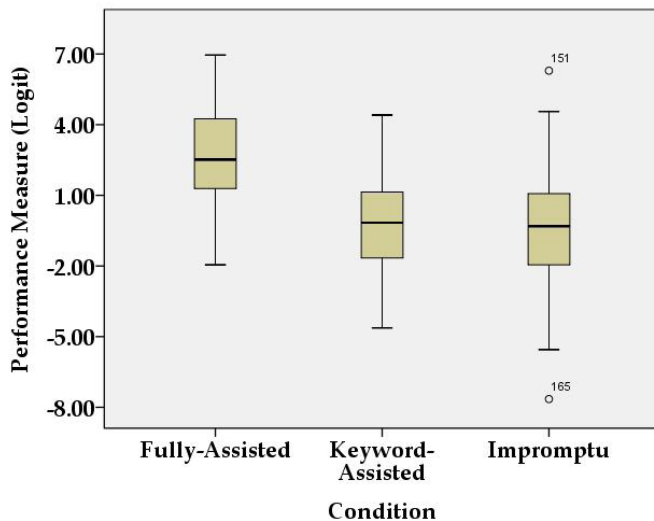


Figure 3. The boxplot of performance measures of three conditions for Group 2.

Further Analysis of the Data

An additional issue was the potential mediating roles played by the keyword list and the length of practice time for performance improvement. Two independent raters checked the total number of words that appeared both on each participant's keyword list and in the actual transcript of the fully assisted and practiced conditions. The keyword-assisted condition was not examined due to the statistically nonsignificant results mentioned above. The rater agreement ratio was 96% for the fully assisted condition and 94% for the practiced condition; and all disagreements were subsequently resolved. The average number of words that participants listed and actually used was 23.64 ($SD = 15.06$) for the fully assisted condition and 24.31 ($SD = 12.07$) for the practiced condition. Average self-reported length of practice was 97.79 minutes ($SD = 72.20$). Both the number of keywords and practice time were positively skewed. Because of the need for normalizing the variables and the power law of learning (Anderson, 2005, p. 188) warranting a strong assumption of a power relationship between practice time or keyword use and gain scores, all variables were log-transformed. Key statistical assumptions (normality, homoscedasticity, nonmulticollinearity) being met, regression analyses were run to test the predictability of the two variables

(time and keywords) for performance gain indicated by the two difference scores: the impromptu/fully assisted scores and the impromptu/practiced scores. Those difference scores were derived by subtracting the measures of the impromptu condition from those of the fully assisted and practiced conditions, which is assumed to reflect gain by practice. Independent variables were the number of keywords along with two covariates: practice time and proficiency (TSST test scores). The results suggested that the number of keywords used, practice time, and proficiency were not statistically significant predictors of either performance difference score, $F(3, 78) = .85, p > .05$, adjusted $r^2 = .00$ for the impromptu/fully assisted and $F(3, 25) = .34, p > .05$, adjusted $r^2 = -.02$ for the impromptu/practiced. These results suggest that the number of keywords actually jotted down and the number of hours of practice did not proportionately improve performance. One might make the criticism that individual variations in keyword use and practice time in the present study could have influenced the effects of conditions. These statistically nonsignificant results of regression analyses address this concern and suggest that the number of keywords used itself does not matter. They also suggest that longer practice does not proportionately boost performance scores; however, the opportunity to practice still seems to have an impact.

Discussion and Conclusion

In the present study I investigated the effect of test practice and keyword use on story-retelling task performance through testing of the TPE and KUE hypotheses. The results were not conclusive. Neither of the competing hypotheses is clearly superior to the other. In Contrast 1 (the fully assisted condition vs. the practiced condition), the KUE hypothesis is supported because the use of keywords further improves performance compared to performance without them. This benefit disappears, however, in the statistically nonsignificant Contrast 6 (keyword-assisted vs. impromptu), suggesting that performance with the keyword list is no better than performance without it. Although the statistically nonsignificant result of Contrast 4 (practiced vs. keyword-assisted) supports neither hypothesis, the results of Contrasts 2, 3, and 5 support the TPE hypothesis. The TPE hypothesis can explain that test practice helps the participants achieve a higher score than does performance without practice (Contrasts 3, 5) by a margin of more than three logits. It also explains the power of practice observed in Contrast 2 (fully assisted vs. keyword-assisted), such that practice, along with the use of a keyword list, leads to a better performance than does a keyword list alone.

The KUE hypothesis was supported in Contrasts 1 and 3 (fully practiced vs. practiced or impromptu), in which the keyword list was used with practice. This suggests that keyword lists have an effect only when used with practice. The test takers in the keyword-assisted condition constructed an immediate keyword list, which emulates pretask planning in previous studies. Thus, this part of the present study seems to replicate the null effects of pretask planning in language testing contexts (e.g., Iwashita, McNamara, & Elder, 2001; Wigglesworth & Elder, 2010).

Although explaining the exact cognitive mechanisms of the differential roles of the keyword list is beyond the scope of the present study, a reasonable speculation is that students at this level (i.e., lower intermediate and upper beginners) could not take advantage of the keyword list without practice. Each condition presented competing demands of cognitive processing to perform the task. In the keyword-assisted condition, the test takers needed to read and understand the passage, plan the retelling, and select and write down the keywords. The keyword list-making task, which should have helped the actual story retelling, might have backfired in the keyword-assisted condition because the test takers were busy making a list rather than planning what to say in the story retelling. In the fully assisted condition, the test takers already understood the story and had practiced for the story retelling and were thus able to save their cognitive resources for the on-line demand of articulating the well-practiced speech to which the keyword list provided cues. In sum, the findings suggest that the keyword list-making task in the keyword-assisted condition may hinder language planning; the dual purposes of the task in anxiety-provoking test contexts limit the availability of cognitive resources for optimal speech performance. This may be the main reason why the keyword list did not help and served only as a security blanket in the keyword-assisted condition. In contrast, the same keyword list facilitated performance in the fully assisted condition. Compared to practice alone (the practiced condition), the test takers had saved their cognitive resources due to practice, thus affording the benefits from the keyword list. A caveat, however, is that this may be true only for learners at this level.

The present study shows that the TPE hypothesis is supported when the test task is identical to the prepared task, which corroborates the bulk of SLA studies on task repetition. Nevertheless, even if the same type of task is used, benefits from practice may not be transferable to different topics, as suggested by the fact that the worst performances occurred in the impromptu condition. One could argue that this is broadly consistent with the

small effects found in test preparation studies, which have indicated that even if the test takers practice using similar types of questions, transferability to the real test is negligible. However, because of the lack of a pretest, it is unclear whether the impromptu performance improved when compared to performance *before* practice. This is one limitation of the present study.

Another limitation is the restriction of sample size and task, which limits the generalizability of the results. In addition, there were limitations regarding research design. The study did not include any analyses using discourse measures, mainly because of their inherent problems in meeting statistical assumptions. This, however, certainly limits the interpretation of the results beyond rated performance.

Finally, it is worth commenting on the role of practice in the classroom performance test. Without the help of practice and the use of a keyword list, the participants in the present study could have demonstrated worse performance, as shown in the impromptu condition. This classroom assessment has thus encouraged the learner to practice to learn, fulfilling its purpose. In the context of a standardized test, test preparation using even part of what is going to be on the test—also called current-form preparation (Popham, 1991)—is claimed to be educationally unjustifiable. However, this assertion needs to be reconsidered in light of performance tests in classroom contexts, because test preparation allows students (a) to feel comfortable by knowing what needs to be practiced; thus, (b) to work hard for practice (positive washback); and (c) to demonstrate their upper limit of performance or the zone of proximal distance in the test. The present study provided at least some support for (b) and (c). For language skills development, preparation for performance tests in the classroom should be used wisely. Further research into the transfer effects of practice in a pre- and posttest design and generalizability of the findings across different proficiency levels will help inform better test preparation for both students and teachers.

Acknowledgements

This study is partially supported by Grants-in-Aid for Scientific Research (Task no. 19520475) awarded to the author. The author is grateful to Dann Gossman, Kunihiro Nagasawa, and reviewers of earlier versions of this paper.

Hidetoshi Saito teaches pre- and in-service EFL teachers at Ibaraki Univer-

sity. His current research interests are in modeling formative assessment and establishing instructional approaches for a three-person discussion format—known as *Interactive English Forums* in Ibaraki Prefecture.

References

- ALC Press (2016). Telephone Standard Speaking Test (TSST). Retrieved from <https://tsst.alc.co.jp>
- Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-Verbal scores. *American Educational Research Journal*, 17, 239-251. <https://doi.org/10.2307/1162485>
- Anderson, J. R. (1999). *Learning and memory: An integrated approach*. New York, NY: John Wiley & Sons.
- Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth.
- Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.
- Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal*, 22, 84-94. <https://doi.org/10.1177/003368829102200106>
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge, UK: Cambridge University Press.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.
- Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23-48). Harlow, UK: Pearson Education.
- Bygate, M., & Samuda, V. (2005). Integrative planning through the use of task-repetition. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 37-74). Amsterdam, the Netherlands: John Benjamins.
- Day, R. R., & Yamanaka, J. (1998). *Impact issues: 30 key issues to help you express yourself in English*. Hong Kong, China: Longman Asia ELT.
- DeKeyser, R. M. (2007). Introduction: Situating the concept of practice. In R. M. DeKeyser (Ed.), *Practice in second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1-18). New York, NY: Cambridge University Press.

- Fisher, Jr., W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 21, 1095. Retrieved from <https://www.rasch.org/rmt/rmt211m.htm>
- Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49, 549-581. <https://doi.org/10.1111/0023-8333.00102>
- Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education*, 14, 75-97. <https://doi.org/10.1080/09695940701272880>
- Gregory, H. (2002). *Public speaking for college and career*. Boston, MA: McGraw-Hill.
- Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly*, 6, 151-167. <https://doi.org/10.1080/15434300902801925>
- Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, 51, 401-436. <https://doi.org/10.1111/0023-8333.00160>
- Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, 19, 357-377. <https://doi.org/10.1093/applin/19.3.357>
- Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 143-164). Amsterdam, the Netherlands: John Benjamins.
- Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perception and practices. *Educational Measurement: Issues and Practice*, 27(2), 28-45. <https://doi.org/10.1111/j.1745-3992.2008.00120.x>
- Linacre, J. M. (n.d.). *Estimation considerations*. Retrieved from <http://www.winsteps.com/facetman/estimationconsiderations.htm>
- Linacre, J. M. (2008). Facets (Version 3.64) [Computer program]. Chicago, IL: MESA Press.
- Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research*, 4, 221-250. <https://doi.org/10.1177/136216880000400303>
- Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. D. Garcia Mayo, M. J. Gutierrez Mangado, & M. M. Adrian (Eds.), *Contemporary approaches to second language acquisition* (pp. 71-91). Amsterdam, the Netherlands: John Benjamins.

- Mackey, A., Kanganas, A. P., & Oliver, R. (2007). Task familiarity and interactional feedback in child ESL classrooms. *TESOL Quarterly*, 41, 285-312.
<https://doi.org/10.1002/j.1545-7249.2007.tb00060.x>
- Muranoi, H. (2007). Output practice in the L2 classroom. In R. M. DeKeyser (Ed.), *Practice in second language: Perspectives from applied linguistics and cognitive psychology* (pp. 51-84). New York, NY: Cambridge University Press.
- Nation, I. S. P., & Newton, J. (2008). *Teaching ESL/EFL listening and speaking*. New York, NY: Routledge.
- Nguyen, T. N. H. (2007). Effects of test preparation on test performance—the case of the IELTS and TOEFL iBT listening tests. *Melbourne Papers in Language Testing*, 12, 1-24.
- Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, 10, 12-15.
<https://doi.org/10.1111/j.1745-3992.1991.tb00211.x>
- Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement*, 22, 121-136.
<https://doi.org/10.1111/j.1745-3984.1985.tb01052.x>
- Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin*, 100, 67-77.
<https://doi.org/10.1037//0033-2909.100.1.67>
- Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*, 12, 24-39.
<https://doi.org/10.1002/j.2333-8504.1993.tb01543.x>
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43, 1-32.
<https://doi.org/10.1515/iral.2005.43.1.1>
- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systematic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39, 369-393.
<https://doi.org/10.1002/tea.10027>
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.
- Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge, UK: Cambridge University Press.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, 7, 1-24. <https://doi.org/10.1080/15434300903031779>

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly*, 26, 693-711. <https://doi.org/10.2307/3586869>

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly*, 10, 196-218. <https://doi.org/10.1080/15434303.2012.721423>

Appendix A

Rating Scale and Main Points of Passages

Table A1. Rating Scale

Level	Score	Grammar & vocabulary	Fluency	Content
Upper	4	<ul style="list-style-type: none"> Makes small errors that never affect comprehensibility Uses complex structures (e.g., subordinate clauses, relative clauses) Uses a wide range of vocabulary and phrases, sophisticated vocabulary 	<ul style="list-style-type: none"> Uses natural pauses Speech flows smoothly and quickly Uses natural fillers for coping with silence (Pronunciation much less influenced by L1) 	<ul style="list-style-type: none"> All main points* are covered
	3	<ul style="list-style-type: none"> Makes occasional small errors, which do not affect overall comprehension Able to produce longer stretches of sentences (complex structures) Uses adequate vocabulary 	<ul style="list-style-type: none"> Speech flows smoothly but contains occasional unnatural pauses Uses a small amount of repetition, pauses 	<ul style="list-style-type: none"> All main points* but one are covered

Level	Score	Grammar & vocabulary	Fluency	Content
	2	<ul style="list-style-type: none"> • Makes numerous small errors that may affect comprehensibility • Sometimes the listener needs sympathy to understand him/her • Uses basic structures • Unable to use rich vocabulary 	<ul style="list-style-type: none"> • Unnatural repetition and reformulation (self-correction) dominates in the speech • Speech contains occasional unnecessary pauses, hesitant • Unnatural intonation and pronunciation • Occasional smoothness may begin to emerge, but not stable • Unable to use appropriate fillers or may use Japanese fillers 	<ul style="list-style-type: none"> • Two points* are missed or not clearly covered
Lower	1	<ul style="list-style-type: none"> • Makes a lot of local errors, even on basic structures, and some global errors that hinder comprehensibility • The listener needs a lot of sympathy for comprehension, but it may not be comprehensible 	<ul style="list-style-type: none"> • Speech contains a lot of pauses including long pauses of 3.0 seconds or more; Frequent repetition, reformulation, breakdown can be observed; Speech is very choppy • Uses unnatural intonation, pronunciation • Has difficulty in continuation • Discards turns • Shows uncertainties in the choice of words • Unable to use fillers or uses Japanese fillers 	<ul style="list-style-type: none"> • All three points* are not clearly covered or missed

Note. *Main points of each passage are described in Table A2.

Table A2. Main Points of Passages

<p>Content: International Relationship</p> <ul style="list-style-type: none"> • Amir & Sachiko are international college students in the U.S., and her boyfriend asks Sachiko to marry him. • Sachiko hasn't yet decided to get married with Amir because she's very confused. • They have several problems such as parents' disapproval, where to live, where to work, and disadvantages that their children might have. • (at least two of these problems should be mentioned)
<p>Content: Doing the Right Thing</p> <ul style="list-style-type: none"> • Yumi's father is dying because of stomach cancer. • Yumi is confused now, and she seeks advice from Dr. Aoki. • They haven't told the truth to her father, because her mother thinks it would shock him and may shorten his life, but Yumi thinks they should tell the truth.
<p>Content: Why Don't You Accept Us?</p> <ul style="list-style-type: none"> • Wing & Jay are a gay couple. • They want to be accepted by society and hope to live an ordinary life. • They have problems in their life because they have to hide the truth from friends, parents, and others.

Appendix B

Readability Indices of the Three Passages

Passages	Dale-Chall readability formula				Fry Grades	Flesch	Lexile	
	Sentence	Word	Diff Wd	D-C score				
Why	30	371	18	5.3	5&6	4	4.3	650L
Doing	30	354	22	5.3	5&6	3	3.3	520L
Int'l	30	353	14	5.2	5&6	4	3.9	510L

Note. Why = “Why Don’t You Accept Us?”; Doing = “Doing the Right Thing”; Int’l = “An International Relationship”; all three passages were taken from Day and Yamanaka (1998). Sentence = the number of sentences; Word = the number of words; DiffWd = the number of difficult words; D-C = Readability scores from the Dale-Chall formula (appropriate for the fourth grade and beyond); Grades = Grade level; Fry = Based on the Fry graph (appropriate for early elementary grades through college); Flesch = the Flesch Grade Level formula (appropriate for upper elementary and secondary levels); Lexile = lexile measures from Meta Metrics Institute.

Appendix C

Summary Statistics of the Rasch Analysis of the All Facets Run

Facets	RMSE	Adj. SE	Separation	Reliability
Raters ¹	.06	0.18	2.72	.88
Participants (performance)	.35	1.20	3.42	.92
Passages	.06	0.35	5.45	.97
Conditions	.08	1.06	13.51	.99
Items	.06	1.19	18.38	1.00

Note. RMSE = root mean square standard error; Adj. SE = adjusted standard error; separation = a measure of the spread of the estimates. ¹Exact agreement among raters was 53.2%. *n* (raters) = 3; *n* (participants) = 82.