Equivalency of Picture-Based Speaking Tasks: An Investigation of Complexity, Accuracy, Lexis, and Fluency

Joe Kakitani
Utsunomiya University / Lancaster University

In experimental research, a pretest–posttest design is often used to examine the effect of treatment or intervention. Establishing equivalency across tests is essential for this type of research to ensure the validity of the study results. However, studies that explore test/task equivalency are scarce in second language research. This study investigates the equivalency of seven picture-based narrative tasks such as those commonly used in second language research and language tests. The oral performances elicited from 20 Japanese university students were analyzed in terms of their complexity, accuracy, lexis, and fluency. Despite controlling for the task-related variables of story length, sequential structure, and storyline complexity, the results were similar only in terms of fluency and not with regard to complexity, accuracy, or lexis. These findings suggest that it is important to determine the equivalency of testing materials as seemingly similar tasks do not necessarily elicit similar oral performances.

実験的研究では、対処や介入の効果を調べるために事前・事後テストデザインがよく用いられる。この種の研究では、結果の妥当性を担保するためにテストの同等性を確立することが重要である。しかし、第二言語研究において、テストやタスクの同等性を調査した研究は極めて少ない。本研究は、第二言語研究や言語テストにおいて頻繁に使用されている写真描写タスクに焦点を当て、7つのタスクの同等性を検証する。そのために、20人の日本人大学生のスピーキングパフォーマンスを、複雑さ、正確さ、語彙、流暢さの尺度を使って分析した。結果は、タスクに関わる要因である物語の長さ、構造、複雑さを統制したにもかかわらず、流暢さにおいてのみタスクの類似性が示され、複雑さ、正確さ、語彙においては差異が認められた。本研究結果は、外見的には似ているタスクであっても同等のスピーキングパフォーマンスに必ずしも結びつかないことを示し、テストに使用する題材の同等性を検証することの重要性を示唆している。

https://doi.org/10.37546/JALTTLT47.2-1

One commonly used experimental design in second language (L2) studies is the pretest–posttest design, which requires demonstration of test equivalency so as not to threaten the validity of the research findings (Mackey & Gass, 2016). Test equivalency is also essential in standardized language testing, where different test forms are supposed to yield consistent results and interpretations. Yet, despite its importance, few published L2 studies to date have explored the topic. This study seeks to address this deficiency in the literature by examining the equivalency of one particular type of testing material, namely picture-based speaking tasks.

Literature Review

Equivalency in Oral Narrative Tasks

Picture-based speaking tasks are commonly used in L2 research (e.g., de Jong & Tillman, 2018; Suzuki et al., 2022; Suzuki & Hanzawa, 2022) and language tests such as EIKEN, one of the leading English-language tests in Japan. Previous studies, though few in number, have explored task equivalency in oral narrative tasks. For instance, Tavakoli and Foster (2011) compared four cartoon stories of varied narrative type and structure by analyzing oral performances based on the linguistic measures of complexity, accuracy, and fluency. Their findings showed that, compared to a one-story narrative, a two-story narrative is likely to result in enhanced complexity because the presence of multiple events requires the speaker to use certain syntactic structures (e.g., subordination). Their findings also suggested that, compared to a loose narrative structure (one in which the order of pictures can be changed), a tight narrative structure (one in which the order of pictures cannot be changed) leads to enhanced accuracy and fluency, because the orderly nature of the narrative events frees up the speakers' attentional resources (Kormos, 2006).

In other studies, de Jong and Vercellotti (2016) and Inoue (2013) examined the equivalency of oral narrative tasks that were a priori deemed similar in terms of sequential structure, storyline complexity, and number of elements. The five tasks chosen in de Jong and Vercellotti's (2016) study produced similar performances in terms of complexity (words per Analysis of Speech Unit [AS-unit]) and accuracy (error-free AS-units) but not with regard to fluency (mean pause length). In fact, one particular task elicited longer pauses, presumably due to lexical retrieval difficulty (i.e., recalling words needed to narrate the story) and task difficulty (i.e., explaining the intention of the character). In a similar vein, In-

oue (2013) compared two narrative tasks that were nearly identical in structure and storyline (i.e., two children playing a trick on their mother in a house). The results were equivalent in terms of fluency (speech rate) but not with regard to complexity (e.g., mean clause length) or accuracy (error-free clauses). The variability was assumed to be due to the varied degree of task complexity, which was manipulated by the change in setting. Namely, Task A depicted scenes both inside and outside the house, whereas Task B showed indoor scenes only. Task A was presumably more cognitively demanding since it required the speaker to think about the connections between the scenes, thus increasing complexity while negatively impacting accuracy of the narrative performance.

The complex findings from these previous studies suggest that it is important to test the equivalency of tasks rather than assume it (Suzuki & Koizumi, 2020). As oral narrative tasks are commonly used in L2 task-based research and language testing, an investigation of their equivalency may provide useful insights for L2 researchers and test developers alike.

Present Study

The current study assessed the equivalency of seven picture-based narrative tasks on measures of complexity, accuracy, lexis, and fluency (CALF). As previous research has shown that various task-related factors can influence L2 oral performance, the current study controlled for task length, structure, and storyline complexity. Its ultimate aim was to identify narrative tasks that are comparable and, thus suitable for a pretest–posttest experimental study. The study was guided by the following research question:

RQ: To what extent do picture-based narrative tasks elicit similar oral performances when task-related variables are controlled?

Method

Participants

The participants were 20 first-language Japanese-speaking 2nd-year university learners of L2 English (12 males, 8 females) between 19 and 20 years of age. Their English proficiency level was approximately the Common European Framework of Reference (CEFR) equivalent of A2–B1, as estimated from their TOEIC scores at the end of their first year of study (M = 556.5, SD = 55.39, minimum = 480, maximum 650).

Materials

The chosen materials comprised seven picture prompts (*Bicycle*, *Race*, *Bus*, *Soccer*, *Picnic*, *Surprise*, and *Hide-and-Seek*) adapted from Heaton (1966, 1975). All prompts were six-frame cartoon stories (except for *Hide-and-Seek*, which had seven frames; however, its sixth and seventh frames were half-size and together depicted one event, making it practically equivalent to the others). All prompts had a tight sequential structure and a storyline that encouraged participants to express the feelings and motivation of the characters (see Appendix A for a link to the materials).

Procedure

Each participant met with the researcher individually online via Microsoft Teams. After giving informed consent, each participant performed the seven monologue tasks. Each participant thus served as his or her own control in the within-subjects design, minimizing error variance (Plonsky & Oswald, 2014). Data collection was divided into two sessions, spanning two consecutive days in order to reduce fatigue. The task order was counterbalanced by randomly assigning the participants to either presentation order A (*Bicycle*, *Race*, and *Bus* on the first day and *Soccer*, *Picnic*, *Surprise*, and *Hide-and-Seek* on the second) or its inverse, order B (*Hide-and-Seek*, *Surprise*, *Picnic*, and *Soccer* on the first day and *Bus*, *Race*, and *Bicycle* on the second). Picture prompts were shown to the students via screen sharing. The students were given 3 minutes of planning time, followed by 4 minutes of speaking time to narrate the story in English. They were told not to take notes or consult a dictionary. Each task included a set of guiding questions in Japanese, which was intended to help clarify the story and give the students additional ideas with regard to content. The guiding questions were only available during the planning time, not during the speaking time. Oral performances were recorded on Microsoft Teams as well as the participant's own mobile device for backup purposes.

Analysis

A total of 140 speech datasets were transcribed and pruned (i.e., excised of filled pauses, repetitions, and self-corrections) based on AS-units. An AS-unit is roughly equivalent to a sentence but can also include commonly found sub-clausal units of speech, such as "Thank you" and "Okay" (Foster et al., 2000). The list of all the indices used in the current analysis can be found in Table 1. Following Norris and Ortega (2009), syntactic complexity was analyzed at three different

levels: sentential, phrasal, and clausal. Accuracy was evaluated using a weighted clause ratio (Foster & Wigglesworth, 2016), assessed by a trained research assistant who is a native speaker of English. Lexical complexity was analyzed using TAALED (Kyle et al., 2021) and TAALES (Kyle et al., 2018). Finally, fluency was analyzed using Praat (Boersma & Weenink, 2018) and a Praat script (de Jong & Wempe, 2009).

For statistical analysis, a series of repeated-measures ANOVAs were performed. The assumption of normal distribution was tested by using histograms and the Shapiro-Wilk tests. When the assumption of normality was violated, a log transformation was performed (Field et al., 2012). In the case of inadequate transformation, the non-parametric counterpart test (i.e., Friedman's ANOVA) was used. The assumption of sphericity was assessed using Mauchly's test, and whenever it was violated, the Greenhouse-Geisser adjustment was used. Significant main effects were further analyzed by performing pairwise comparisons with Bonferroni correction.

Table 1

List of Measures

Syntactic complexity
1. *Mean length of AS-unit.* Average number of words per AS-unit
2. *Mean length of clause.* Average number of words per clause
3. *Clauses per AS-unit.* Average number of clauses per AS-unit
Accuracy
4. *Weighted clause ratio.* Total clause accuracy score divided by total number of clauses
Lexical complexity
5. *Measure of textual lexical diversity.* Mean length of sequential word strings that maintains a given type-token ratio value
6. *Word frequency.* Average logarithmic frequency of content words based on SUBTLEXus
7. *Word familiarity.* Average familiarity score of content words based on MRC Psycholinguistic Database
8. *Word imageability.* Average imageability score of content words based on MRC Psycholinguistic Database
Fluency
9. *Articulation rate.* Mean number of syllables per second, excluding the duration of pauses
10. *Speech rate.* Mean number of syllables per second, including the duration of pauses
11. *Mean length of fluent run.* Mean number of syllables produced in utterances between pauses (.25 seconds and above)

Results

Syntactic Complexity

The effect of task was statistically significant for the mean length of AS-unit (MLAS), $F(6, 114) = 5.93$, $p < .001$, $\eta^2 = .16$ (see Table 2). The results of the post hoc comparisons are shown in Appendix B. *Bus* elicited a longer MLAS compared to *Bicycle*, *Soccer*, and *Hide-and-Seek*. MLAS was also longer for *Picnic* than *Hide-and-Seek*. The effect of task was also statistically significant for the mean length of clause (MLC), $F(6, 114) = 11.71$, $p < .001$, $\eta^2 = .314$. *Bus* produced a longer MLC than all the other tasks. MLC for *Hide-and-Seek* was in turn shorter than for *Race*, *Soccer*, and *Surprise*. Finally, the effect of task for the mean number of clauses per AS-unit was not statistically significant, $\chi^2(6) = 5.89$, $p = .436$.

Accuracy

The effect of task for the weighted clause ratio (WCR) was statistically significant, $F(6, 114) = 7.94$, $p < .001$, $\eta^2 = .242$ (Table 3). Particularly, *Hide-and-Seek* produced a higher WCR than all the other tasks except for *Soccer* (see Appendix C).

Lexical Complexity

Four measures were used to assess lexical complexity (see Table 4). The measure of textual lexical diversity (MTLD) shows the range of words used in a text, with a higher score indicating a higher diversity (McCarthy & Jarvis, 2010). Word frequency, word familiarity, and word imageability provide word information scores based on large corpora of texts. For these three measures, lower scores indicate the use of more sophisticated words (for more detailed information, see Kyle & Crossley, 2015). The effect of task was statistically significant for MTLD, $F(6, 114) = 4.91$, $p < .001$, $\eta^2 = .164$. *Hide-and-Seek* elicited greater MTLD than did *Bicycle* and *Bus*. The effect of task was also statistically significant for word frequency, $F(6, 114) = 9.12$, $p < .001$, $\eta^2 = .266$. *Race* elicited words with lower word frequency scores compared to *Bicycle*, *Bus*, *Soccer*, and *Surprise*.

Furthermore, word frequency was lower for *Picnic* than for *Bus*. Word familiarity was also statistically different across tasks, $F(6, 114) = 14.61$, $p < .001$, $\eta^2 = .396$. *Bicycle* elicited higher word familiarity scores compared to all the other tasks except *Surprise*. Word familiarity was also higher for *Bus* than for *Race* and *Soccer*. *Picnic* and *Surprise* also both resulted in higher word familiarity compared to *Soccer*. Finally, the effect of task was also statistically significant for word imageability, $F(6, 114) = 13.93$, $p < .001$, $\eta^2 = .325$. All tasks except *Bus* produced higher word imageability scores than did *Race*. In turn, all tasks except *Race* showed higher word imageability than did *Bus* (see Appendix D).

Fluency

Finally, fluency was evaluated by articulation rate, speech rate, and mean length of fluent run—measures commonly used in L2 fluency research (e.g., de Jong & Perfetti, 2011; Suzuki & Kormos, 2020). The effect of task was not statistically significant for any of the three measures: $F(2.62, 49.82) = 2.58$, $p = .071$, $\eta^2 = .032$, and $F(6, 114) = 0.96$, $p = .455$, $\eta^2 = .014$, and $F(2.86, 54.38) = 1.69$, $p = .181$, $\eta^2 = .03$, respectively (see Table 5). All tasks, therefore, elicited comparable performances in terms of fluency.

Table 2
Descriptive Statistics and ANOVA (or Friedman's ANOVA) Results for Syntactic Complexity Measures

	MLAS		MLC		CPAS	
	M	*SD*	*M*	*SD*	*M*	*SD*
Bicycle	7.55	1.41	5.64	0.59	1.35	0.26
Race	8.08	1.15	5.96	0.73	1.36	0.19
Bus	8.90	1.60	6.70	0.77	1.35	0.31
Soccer	7.61	0.98	5.72	0.62	1.34	0.17
Picnic	8.13	1.30	5.79	0.80	1.41	0.23
Surprise	7.86	1.01	5.93	0.68	1.34	0.21
Hide-and-Seek	7.13	1.09	5.10	0.53	1.41	0.24
p	<.001		<.001		.436	

Note. MLAS = mean length of AS-unit; MLC = mean length of clause; CPAS = clauses per AS-unit.

Table 3
Descriptive Statistics and ANOVA Results for Accuracy Measure

	WCR	
	M	*SD*
Bicycle	0.65	0.09
Race	0.70	0.08
Bus	0.69	0.08
Soccer	0.73	0.09
Picnic	0.73	0.08
Surprise	0.70	0.08
Hide-and-Seek	0.80	0.05
p	<.001	

Note. WCR = weighted clause ratio.

Table 4
Descriptive Statistics and ANOVA Results for Lexical Complexity Measures

	MTLD		Word Frequency		Word Familiarity		Word Imageability	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Bicycle	22.21	4.53	4.03	0.19	599.58	5.65	469.80	20.56
Race	25.02	3.70	3.78	0.29	581.94	7.90	436.87	16.94
Bus	22.12	5.18	4.13	0.11	591.08	5.43	440.86	23.06
Soccer	25.67	6.24	4.09	0.15	582.34	4.90	470.11	22.12
Picnic	26.93	6.21	3.98	0.13	589.84	8.21	477.40	28.36
Surprise	25.28	7.01	4.13	0.16	591.53	8.84	462.57	18.90
Hide-and-Seek	30.01	7.09	4.02	0.18	585.91	8.40	471.63	21.33
p	<.001		<.001		<.001		<.001	

Note. MTLD = measure of textual lexical diversity.

Table 5

Descriptive Statistics and ANOVA Results for Fluency Measures

	AR		SR		FR	
	M	*SD*	*M*	*SD*	*M*	*SD*
Bicycle	2.74	0.54	1.17	0.28	2.64	0.69
Race	2.64	0.65	1.18	0.24	2.56	0.60
Bus	2.68	0.50	1.24	0.28	2.92	0.89
Soccer	2.88	0.41	1.19	0.28	2.60	0.80
Picnic	2.89	0.42	1.23	0.34	2.78	0.91
Surprise	2.76	0.41	1.14	0.31	2.51	0.64
Hide-and-Seek	2.77	0.48	1.20	0.35	2.60	0.97
p	.071		.455		.181	

Note. AR = articulation rate; SR = speech rate; FR = mean length of fluent run.

Discussion

The current study investigated the equivalency of picture-based narrative tasks using the CALF framework. Oral performances were indeed similar in terms of fluency. However, substantial differences were found with respect to complexity, accuracy, and lexis. The *Bus* task overall elicited more syntactically complex narrative performances. One possible explanation for this result might be that this particular task requires reference to a previous event in the story—the second bus passes the first bus, which the boys could not ride earlier—leading to longer utterances. However, this explanation is inadequate considering that other tasks (e.g., *Bicycle*) have a similar storyline. Thus, another possible reason might be that there are simply more details in the *Bus* task (e.g., road condition, bus numbers, clocks showing the time) that prompt longer utterances. The *Hide-and-Seek* task, by contrast, elicited relatively less complex performances (i.e., shorter AS-units and clauses). This result may be due to the fact that most of the events take place in the foreground (Tavakoli & Foster, 2011), helping the speakers to narrate the story in a straightforward manner (e.g., without subordination).

In terms of accuracy, the *Hide-and-Seek* task produced higher accuracy scores on average. Considering that this task elicited relatively less syntactically complex utterances, it is possible that speakers were able to pay more attention to accuracy because of the simplicity of the task (Skehan & Foster, 2001). However, another possible, and perhaps more plausible, explanation might be that L2 learners simply have fewer chances of making mistakes in shorter utterances. Although a weighted clause ratio (WCR) allows for a more fine-grained assessment of accuracy than do global indices (e.g., error-free AS-units), the rating system is still subject to the influence of clause length. Indeed, WCR was negatively correlated with the mean length of clause, $r = -.28$, $p < .001$. The findings thus suggest that accuracy and complexity should be interpreted in tandem to draw a nuanced conclusion about task equivalency.

With regard to lexical diversity, the *Hide-and-Seek* task produced relatively higher MTLD. This result could be due to the fact that the story involves many characters and objects (e.g., boys, girls, a statue, a vase) and actions (e.g., hiding, falling, breaking, coming out). Thus, the number of elements may be a factor that can significantly influence the lexical diversity of a narrative performance. The measures of word frequency, word familiarity, and word imageability also showed substantial variability across tasks. Unlike lexical diversity, these indices provide word information scores based on large text corpora (Kyle & Crossley, 2015). The observed variability could be ascribed to the nature of closed tasks (Pallotti, 2009). In closed tasks, the content of speech is predefined for the most part, as the given prompt necessitates the use of certain expressions to complete them. For instance, the *Bicycle* task requires the speaker to use words such as *bicycle*, *road*, and *car*, all of which the MRC database designates as highly familiar words. The considerable variability across tasks in terms of lexical complexity suggests that the content words elicited in a narrative task might be determined by task design features and the semantic content that L2 learners need to express (see de Jong & Vercellotti, 2016, for similar discussions).

Finally, the seven tasks elicited similar performances in terms of fluency. There are two possible reasons for this. First, the tasks all had a tight sequential structure. In line with previous research (Inoue, 2013; Tavakoli & Foster, 2011), the chronological sequence of the narrative story in the present study probably helped the speakers to narrate each story with relative ease. Second, the materials used in the current study were taken from a single source (i.e., a single author), controlling for the aesthetic quality of the prompts. Different cartoon artists have different drawing styles, and these differences in artistic touch could potentially lead to varying degrees of difficulty for interpreting the intentions and emotions of the characters. As a case in point, in de Jong and Vercellotti's study (2016), the three cartoon prompts drawn by a single illustrator elicited similar performances, while the other two prompts drawn by different artists led to

statistically different fluency results. The current findings suggest that the aesthetic aspect of materials should also be taken into account when comparing and administering picture-based narrative tasks.

Conclusion and Implications

Despite controlling for task-related factors, the oral performances elicited by seven narrative tasks showed substantial differences in terms of complexity, accuracy, and lexis. The current findings bear important implications for L2 researchers, test developers, and instructors. First, as the chosen tasks elicited similar performances in terms of fluency, it is probably appropriate to use any mixture of these tasks for testing purposes in a pretest–posttest study investigating L2 speakers' fluency development. However, because differences were found in terms of syntactic complexity, accuracy, and lexical complexity, caution needs to be taken if research involves analyses of these measures. To minimize potential task effects, it is recommended that researchers counterbalance the order of test materials (Suzuki & Koizumi, 2020). In many testing programs, parallel tasks are used with the assumption that they elicit similar performance from test takers. However, the current findings suggest that this might not be the case. Rather than assuming task equivalency, the test developers should consider piloting their tasks to establish true comparability of the results and their interpretations across different test forms. Finally, from a pedagogical perspective, it may be plausible to use relatively easier tasks first and move on to more difficult tasks, taking the L2 learners' developmental processes into consideration (Lambert & Kormos, 2014). Examining the relative difficulty of instructional tasks may thus provide L2 instructors with a basis for making more informed decisions regarding classroom practice (e.g., L2 fluency training using task repetition). As the scope of the current study was limited to oral narrative tasks, future research should explore the equivalency of other types of tasks as well.

References

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. http://www.praat.org

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

de Jong, N., & Tillman, P. (2018). Grammatical structures and oral fluency in immediate task repetition. In M. Bygate (Ed.), *Learning language through task repetition* (pp. 43–73). John Benjamins.

de Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 387–404. https://doi.org/10.1177/1362168815606161

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116. https://doi.org/10.1017/S0267190515000082

Heaton, J. B. (1966). *Composition through pictures*. Longman.

Heaton, J. B. (1975). *Beginning composition through pictures*. Longman.

Inoue, C. (2013). *Task equivalence in speaking tests: Investigating the difficulty of two spoken narrative tasks*. Peter Lang.

Kormos, J. (2006). *Speech production and second language acquisition.* Routledge.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, *50*, 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 Research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607–614. https://doi.org/10.1093/applin/amu047

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). Routledge.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Plonsky, L., & Oswald, F. L. (2014). How big is "Big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 183–205). Cambridge University Press.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, Y., Eguchi, M., & de Jong, N. (2022). Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective. *TESOL Quarterly*, 1–30. https://doi.org/10.1002/tesq.3103

Suzuki, Y., & Hanzawa, K. (2022). Massed task repetition is a double-edged sword for fluency development: An EFL classroom study. *Studies in Second Language Acquisition*, *44*(2), 536–561. https://doi.org/10.1017/S0272263121000358

Suzuki, Y., & Koizumi, R. (2020). Using equivalent test forms in SLA pretest-posttest design research. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 457–467). Routledge.

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *61*, 37–72. https://doi.org/10.1111/j.1467-9922.2011.00642.x

Joe Kakitani is an assistant professor at Utsunomiya University and a PhD candidate at Lancaster University. He holds a master's degree in TESOL from the University of Southern California. His research interests include L2 speech fluency, speech production, distributed practice, and language learning psychology.

Appendix A

A link to the materials: https://osf.io/73drq/?view_only=01e103a1d8e145c4b353d17b49b2ac44

Appendix B

Pairwise Comparison Results with Bonferroni Correction for Complexity Measures

		Effect size (*d*)	*p*
MLAS	*Bus > Bicycle*	1.16	<.001
	Bus > Soccer	0.79	.026
	Bus > Hide-and-Seek	1.00	.002
	Picnic > Hide-and-Seek	0.72	.042
MLC	*Bus > Bicycle*	1.22	.001
	Bus > Race	1.04	.001
	Bus > Soccer	1.12	.002
	Bus > Picnic	1.03	.004
	Bus > Surprise	0.82	.041
	Bus > Hide-and-Seek	2.62	<.001
	Race > Hide-and-Seek	1.14	.002
	Soccer > Hide-and-Seek	0.93	.025
	Surprise > Hide-and-Seek	1.10	.002

Note. MLAS = mean length of AS-unit; MLC = mean length of clause.

Appendix C

Pairwise Comparison Results with Bonferroni Correction for Accuracy Measure

		Effect size (*d*)	*p*
WCR	*Hide-and-Seek > Bicycle*	1.63	<.001
	Hide-and-Seek > Race	0.97	.010
	Hide-and-Seek > Bus	1.39	<.001
	Hide-and-Seek > Picnic	0.84	.045
	Hide-and-Seek > Surprise	1.20	<.001

Note. WCR = weighted clause ratio.

Appendix D

Pairwise Comparison Results with Bonferroni Correction for Lexical Complexity Measures

		Effect size (*d*)	*p*
MTLD	*Hide-and-Seek > Bicycle*	0.81	.036
	Hide-and-Seek > Bus	1.15	.002
Word Frequency	*Bicycle > Race*	0.85	.039
	Bus > Race	1.15	.001
	Soccer > Race	1.14	.002
	Surprise > Race	1.09	.003
	Bus > Picnic	1.07	.002
Word Familiarity	*Bicycle > Race*	2.07	<.001
	Bicycle > Bus	1.18	<.001
	Bicycle > Soccer	2.66	<.001
	Bicycle > Picnic	1.08	.002
	Bicycle > Hide-and-Seek	1.21	<.001
	Bus > Race	0.90	.011
	Bus > Soccer	1.07	.002
	Picnic > Soccer	0.91	.019
	Surprise > Soccer	0.79	.038
Word Imageability	*Bicycle > Race*	1.27	<.001
	Soccer > Race	1.36	<.001
	Picnic > Race	1.28	<.001
	Surprise > Race	1.32	<.001
	Hide-and-Seek > Race	1.28	<.001
	Bicycle > Bus	0.97	.001
	Soccer > Bus	1.06	.002
	Picnic > Bus	1.22	<.001
	Surprise > Bus	1.07	.012
	Hide-and-Seek > Bus	1.22	<.001

Note. MTLD = measure of textual lexical diversity.

Win a Complimentary TESOL Membership!

JALT's International Affairs Committee (IAC) is pleased to announce that seven (7) lucky JALT members will win one-year complimentary memberships to the TESOL International Association! You can see what kind of benefits are offered for TESOL members here: https://www.tesol.org/about-tesol/membership/membership-benefits

Participants must be active and current JALT members, and not a TESOL member (or someone who has not been a TESOL member in the past five years). To sign up for your chance to win, please sign up at:

https://forms.gle/PBY6mjMg48j8Qb6i8

The deadline to sign up is April 23rd, 2023, and seven names will be chosen at random from the pool of eligible applicants. The seven lucky winners will be announced this May. Good luck!