Talking about oneself

Effects of personalized prompts on L2 learners' speech

Charles M. Mueller and Peter Richardson
Fuji Women's University, Japan | Hokkaido Bunkyo University, Japan

Second language instructors often have students talk about their own experiences rather than abstract impersonal topics. Intuitively, such topics seem more likely to encourage student engagement. Unfortunately, virtually no empirical research has examined the effects of personal prompts on spoken output. To address this gap in research, the current study ($N = 117$) compares the spoken output of Japanese university English students who responded to a personal prompt with students responding to an impersonal prompt. Output was recorded in transcripts and then analyzed using a battery of measures related to complexity, accuracy, and fluency. Findings showed that personalized prompts were associated with greater fluency. Moreover, there was some evidence that impersonal prompts led to output with greater lexical complexity. Rates of accuracy were similar in both groups. Correlational analysis suggested that lexical sophistication was associated with reduced fluency. The conclusion addresses practical implications and avenues for further research.

Keywords: speaking, prompts, self-reference, fluency, accuracy, complexity

1. Introduction

A key concern in second language instruction is the creation of tasks and activities that promote the development of speaking skills. Good speaking tasks have several characteristics. On the one hand, they are engaging and encourage students to speak. At the same time, if they are sufficiently challenging, they push students to produce language that is sophisticated, accurate, and appropriate for a given situation (Swain & Lapkin, 1995). A key set of task variables involves the nature of the speaking prompt.

https://doi.org/10.1075/jsls.21023.mue | Published online: 31 December 2021
Journal of Second Language Studies 5:2 (2022), pp. 344–363. ISSN 2542-3835 | E-ISSN 2542-3843

This paper reports the results of an experiment in which speaking prompts were manipulated. The paper begins with a review of relevant research, with a focus on the self-reference effect, autobiographical memory, and potential interactions of these systems with speech processing. This is followed by a report of an experiment that examined the effect of personalized prompts on the complexity, accuracy, and fluency of the spoken output produced by Japanese university students learning English.

1.1 Literature review

Within education, personalized activities are often promoted on both cognitive and affective grounds. From a cognitive perspective, these activities are thought to promote "more meaningful learning", "more active processing of incoming information", and better transfer of learning from pedagogical tasks to assessed performance (Mayer et al., 2004, p. 391). Since typical conversations involve the expression of a speaker's thoughts, experience, and feelings, speakers presumably find it natural to talk about themselves. Such tasks are thus thought to facilitate both the production and comprehension of messages (cf. Moreno & Mayer, 2000), and are also more likely to be regarded as relevant, and thus more motivating, to learners (Chang & Lehman, 2002; Keller, 1987).

Personalized prompts also receive potential justification from research on the self-reference effect. Empirical studies have shown that material associated with the self is more readily retrieved during recall and recognition tasks (Bellezza, 1984; Symons & Johnson, 1997). The mechanisms mediating the self-reference effect on memory have been debated, but the effect is thought to be underscored by a range of factors to include elaboration (Klein & Loftus, 1988), organization (Klein & Kihlstrom, 1986), and evaluation (Ferguson et al., 1983). In a seminal experiment, Rogers et al. (1977) examined the effect of self-reference on participants' memory of trait adjectives. Participants were asked to judge whether some of these adjectives described themselves. Later, items presented in the self-reference condition were better recalled than other items.

Advantages for personalized prompts can also be predicted from the organization and functions of autobiographical memory (AM). Research on AM suggests that the self provides an extensive framework of internal cues, which are organized in the form of rich, coherent, and highly accessible autobiographical memories (Conway, 2005; Conway & Loveday, 2015; Thomsen, 2015). While most memories that we form throughout our daily lives are on a steep forgetting trajectory, a small subset of memories is more readily retrieved due to associations with the self. Thus, the self is said to constitute "one of the richest, most elaborate knowledge structures in memory" (Kihlstrom et al., 1988, p. 150). While self-

reference effects are often discussed in relationship to episodic memory, some evidence suggests that the effects extend to semantic memory as well. Klein (2012) provides some evidence for this in his discussion of several patients with neural damage. He cites the case of "D. B.", an elderly patient who was profoundly amnesic as a result of anoxia following a heart attack. In spite of the patient's severe amnesia and inability to recall previous actions and experiences from his entire life or even from minutes earlier, he was able to make consistent statements about his own personality traits. These statements were also consistent with descriptions of his personality by others who knew him (e.g., his daughter). Such findings suggest that even in the case of semantic knowledge, there is a dissociation between knowledge of the self and other knowledge, with the former much less vulnerable to forgetting.

What relevance could the self-reference effect and autobiographical memory have for second language acquisition and for L2 speaking skills in particular? One possible answer comes from Levelt's (1989) speaking model. Levelt partitions the various processes involved in fluent speech production into three processing components, with each component producing the output for the downstream component. The first component, the Conceptualizer, conceives of an intention and thus selects relevant information to be expressed. This component is therefore responsible for message generation as well as for monitoring activities that keep track of relevant information such as the content of the previous utterance. A crucial point, highly relevant to the current discussion, is that these processes require the speaker's constant attention. In other words, they tend to deplete working memory resources that could be allotted to other tasks (Kerz & Wiechmann, 2020; Wu & Ma, 2016). The preverbal message created in the Conceptualizer then feeds into the Formulator where it receives grammatical and phonological encoding. The processing within the Formulator is normally rapid and automatized in native speakers. Non-native speakers, on the other hand, may rely more heavily on working memory resources as they manipulate grammatical constructions and select appropriate lexical items to fill the slots within these constructions. Based on the processing with the Formulator, a phonetic plan is generated that feeds into the Articulator, which is responsible for the motor movements of actual speech.

L2 learners who are asked to speak on personalized topics can presumably draw on the rich database of knowledge related to the self. If this assumption is on the right track, it can be furthermore predicted that in the case of self-reference tasks, attentional resources that are typically allotted to the Conceptualizer can be freed up for other tasks taking place in the Formulator such as the facilitation of grammatical encoding or searches for appropriate lexemes. Put simply, L2 learners asked to talk about themselves will presumably spend less time thinking about

what to say, and thus gain more time to think about *how* to express their message (cf. Bygate, 2001).

Despite the critical role of prompts in organizing speaking tasks, no studies have directly examined the effects of personalization of prompts on L2 spoken output. However, one recent study by Mueller and Kraus (2018) did examine the effect of personalized prompts on *written* output. The study reported on a within-subjects comparison of 36 Japanese-L1 female EFL participants who wrote timed essays on one personal and one impersonal topic. Variables related to personalization, the topic, and presentation order were both counter-balanced across participants. The essays were then assessed in terms of complexity, accuracy, and fluency. The study found that personalized prompts led to significantly greater accuracy and more fluency, operationalized as the number of words produced (cf. Mueller, 2010, Exp. 2). As assessed through mean length of T-unit (MLTU), essays in response to impersonal prompts exhibited a trend toward greater complexity ($p = .02$). It should be noted that a T-unit (Hunt, 1970) is defined as a dominant clause and its dependent clauses, and it is often a sentence.

When assessing the quality of L2 speaking, researchers have generally focused on complexity, accuracy, and fluency (Bui & Skehan, 2018; Housen & Kuiken, 2009; Housen et al., 2012; Skehan, 1998, 2009; Wright, 2020). Greater complexity and accuracy are associated with a more sophisticated L2 knowledge system and the ability to access knowledge that is less proceduralized. Greater fluency is thought to reflect greater control and automatization of the L2 knowledge base (Michel, 2017). In light of the conclusions so far, memory research on the self-reference effect and Levelt's production model would suggest that tasks involving self-reference should be easier than similar tasks that do not involve self-reference.

1.2 Research questions and hypotheses

Based on Skehan and Foster (1997) and the memory research discussed above, it will be hypothesized that self-reference speaking prompts increase the amount of learner resources available for focus on form and thus lead to greater fluency. Predictions regarding complexity and accuracy are less straightforward. Skehan's (1998) Trade-Off Hypothesis claims that increased task demands degrade fluency, accuracy, and complexity. Robinson (2005), on the other hand, claims that tasks can be difficult along both resource-directing and resource-dispersing dimensions. Resource-directing dimensions are described as "those in which the demands on language use made by increases in task complexity, and the increased conceptual demands they implicate, can be met by specific aspects of the linguistic system" (p. 4). Robinson claims that increased complexity along resource-directing dimensions has the potential to direct "learners' attentional and mem-

ory resources to the way the L2 structures and codes concepts" (p. 4). Resource-dispersing dimensions, in contrast, do not direct learners' attention to aspects of the linguistic code. Robinson claims that tasks that increase complexity along resource-dispersing dimensions (e.g., speech in which there is no planning time) are, nevertheless, valuable, as they simulate typical situations in which language is used.

The Mueller and Kraus (2018) study is the only empirical study to examine the effect of personalized prompts on complexity, accuracy, and fluency of output. The study's participants exhibited greater fluency and accuracy in the personalized prompt condition and a trend toward greater complexity in the impersonal prompt condition. It is therefore assumed that personalization of speaking prompts makes speaking tasks easier and thus leads to greater fluency and accuracy, whereas impersonal prompts make speaking tasks more difficult along resource-directing dimensions and thus lead to greater complexity. The current study thus investigated three research questions:

R1: Which prompt type, personal or impersonal, leads to greater complexity in the spoken output of L2 learners?
R2: Which prompt type, personal or impersonal, leads to greater accuracy in the spoken output of L2 learners?
R3: Which prompt type, personal or impersonal, leads to greater fluency in L2 speech?

These research questions are accompanied by the following three hypotheses, which are based on theoretical considerations and previous research findings:

H1: Japanese-L1 learners of English who are responding to an impersonal prompt will produce more complex output than similar learners who are responding to a personal prompt.
H2: Japanese-L1 learners of English who are responding to a personalized prompt will produce more accurate output than similar learners who are responding to an impersonal prompt.
H3: Japanese-L1 learners of English who are responding to a personalized prompt will exhibit greater fluency than similar learners who are responding to an impersonal prompt.

2. Method

2.1 Participants

The study's participants ($N = 117$) were first year students attending a large public university in Japan. Their ages ranged from 19 to 20 years old, and their language levels ranged from B1 to B2 on the Common European Framework of Reference for Languages (Council of Europe, 2001). The students were all members of a first-year, compulsory integrated skills class that consisted of the following mix of majors: Nursing, Pharmacology, Physiotherapy, Engineering, Veterinary Medicine, Law, Law and Politics, Fisheries, Medicine, and Economics, in addition to students who were yet to decide their major. Out of the 117 participants, 38 were female.

2.2 Procedure

One-on-one interviews were conducted with participants using Zoom. To ensure that participants' preparation for the interviews did not affect results, participants were informed that the graded portion of the interviews would be focused on their response to how the COVID-19 crisis had changed their lives, and that this would be preceded by an ungraded warm-up phase in which other topics would be discussed. The experimental task occurred during this "warm-up" phase. The data elicitation was preceded by some small talk to ensure that the students were comfortable and at ease. Participants were then informed that there would be a warm-up topic and that they would have one minute to respond to the topic. The participants in the personal prompt condition were asked to draw on their personal experiences, whereas those in the impersonal prompt condition were asked to think about Japanese students in general. Participants were randomly assigned to one of the two conditions (i.e., personal or impersonal prompt). The prompts were displayed on the screen in both English and Japanese to minimize comprehension issues. The personal prompt said:

> *Describe how you study English and the difficulties you experience in your English studies. Draw on your personal experiences to answer this question.*

The impersonal prompt said:

> *Describe how Japanese students in general study English and the difficulties they experience in their English studies. Do not think about your own personal experience.*

The entire conversation (both visual and audio elements) was recorded using the "record" function on Zoom. The segment used for analysis was taken from the first minute from the point at which the participant began talking. This segment was transcribed, and the transcription was then checked by both authors.

2.3 Analysis

When multiple statistical tests are conducted within a single experiment, experiment-wise error is an issue. While this can be accounted for through various correction procedures (e.g., Bonferroni adjustments), this results in more stringent alpha values and the increasing likelihood of Type 2 error. To avoid experiment-wise error while simultaneously avoiding the use of overly stringent alpha levels, the three hypotheses in the current experiment were each tested using a single measure. All other measures were regarded as exploratory.

Participants' speaking transcripts were analyzed to determine their complexity, accuracy, and fluency. According to Michel (2017), complexity refers to "the size, elaborateness, richness, and diversity of the L2 performance" (p. 50). A simple method for measuring complexity of linguistic production is to measure the length of sentences, since longer sentences tend to be constructed with complex syntactic units, such as subordinate clauses. A pitfall of this approach is that writers and speakers will often produce run-on sentences with the word *and*, which result in inflated measures of sentence length. These measures lack construct validity since they do not reflect greater morphosyntactic complexity. To overcome this issue, researchers have generally examined production in terms of the mean length of the T-unit (Gaies, 1980; Hunt, 1970). This type of analysis provides rules for dividing run-on sentences to prevent the inflation of complexity measures.

While fairly easy to apply to written data, the T-unit often proves difficult when working with oral data due to speech's more fluid nature. For this reason, Foster et al. (2000) have developed a more elaborated definition of the measure, which they call the Analysis of Speech Unit (AS-unit). They define this unit as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (p. 365). They provide clear criteria for dealing with issues that are specifically relevant to speech, such as how to treat disfluency features such as false starts and self-corrections. In the analysis used in this paper, the speech data were divided into AS-units using their criteria, and each unit was then further divided into clauses. The mean length of an AS-unit was calculated based on the number of tokens per unit.

In addition to AS-unit calculations, which were used to test H1, a further battery of measures was used to explore the complexity of participants' speech, to

include (1) the type-token ratio, (2) the percentage of words from the first 1000 most frequent word families in English, (3) the average frequency of each token, (4) the average frequency of each type, and (5) the average number of syllables per word. These measures were calculated using the Text Inspector (2018) website. The type-token ratio shows the number of unique words used by a participant. It is calculated by dividing the number of tokens (i.e., the total number of running words) by the number of types (i.e., unique words) in a text. A type-token ration of 1.0 indicates that no words were repeated, whereas low numbers closer to zero indicate extensive repetition. High type-token ratios are considered an indication of lexical sophistication.

Another indication of lexical sophistication is decreased use of words that are within the 1000 most common words in English. In addition to providing counts for this list, the Text Inspector website also calculates the frequency at which each word (as either a token or a type) in a transcript occurs per million words in the COCA (a large American corpus). The website then averages these frequencies for all the words in a submitted text sample. For example, the English word *study* occurs frequently in English (263 times per million words), whereas the word *empirical* is rare (occurring 12 times per million words). Lexical sophistication can thus be operationalized as use of words that occur, on average, less frequently in typical English discourse.

Finally, the number of syllables per word provides a useful measure of the length of words in a text. Words with more syllables tend to be less frequent, more precise in meaning, morphologically complex, and are often Latinate. They are thus more closely associated with academic and formal registers. It should be noted that the five additional measures of complexity discussed here are not all independent but are instead highly correlated. They have been included to provide a more fine-grained analysis of the speech transcripts and to facilitate comparisons with studies that adopt varied measures of complexity.

Accuracy has been defined as "the target-like and error-free use of language" (Michel, 2017, p. 50). It has been operationalized in a number of ways, perhaps most straightforwardly as the number of error-free units (e.g., sentences, T-units, or AS-units). One weakness of such measures is that they are insensitive to the relative severity of an error. For example, a long unit missing an indefinite article ends up treated as equivalent to a unit with errors so egregious that it is unintelligible. To better account for the relative severity of errors, the current study adopted Foster and Wiggleworth's (2016) weighted clause ratio, according to which, the accuracy of a clause is given a score based on the severity of the errors (pp. 106, 107) as shown in Table 1.

Fluency has been operationalized in terms of three subdimensions related to speed, breakdown, and repair (Tavakoli & Skehan, 2005). Another option is

Table 1. Weighted Clause Ratio Used to Assess Accuracy

Category	Points	Definition
Accurate	1.0	The clause is accurately constructed.
Level 1	0.8	The clause has minor errors that do not compromise meaning.
Level 2	0.5	The clause contains serious errors but meaning is recoverable.
Level 3	0.1	The clause has very serious errors with intended meaning far from obvious.

to use a measure of syllables per minute. One concern with this latter option in the current study was that Japanese speakers, obeying phonotactic constraints in their L1 that forbid non-nasal consonants in the coda position, often add a paragogic vowel after final consonants when they speak English. When the vowel is shortened by a speaker's attempt to approximate the English target sound, it can become difficult for coders to determine whether the truncated vowel constitutes an additional syllable. For this reason, words per minute was selected as a more appropriate measure for the current study. Silences, repetitions, and self-corrections serve as indications of disfluency. In terms of cognitive processing, speed indicates control and access to proceduralized and automatized knowledge; breakdown reflects problems in stages related to planning and conceptualizing the message; and repair reflects the intervention of monitoring processes (Michel, 2017).

To better show how the analysis was conducted, we presented below the transcript of one of the interviews, which was conducted with a participant in the personalized prompt condition:

> I start to study English :: {since} when I was twelve years old. |
> {I go to the}
> I go to {the} my friend's house. |
> my friend's mother is from Australia ::
> so she teaches me English |and
> {I taught I I study I taught} I was taught her {from in} |
> I studied English :: by listening {English} English song ::
> and talked too.

As can be seen, the AS-units have been separated by a vertical bar. "My friend's mother is from Australia, so she teaches me English" and "I was taught her" are independent clauses separated by *and*, so they are regarded as separate AS-units. The entire excerpt thus consists of five AS-units. Disfluencies such as false starts have been enclosed in curly brackets. In the transcript, sub-clausal units are marked with doubled full colons. Excluding the words marked as disfluencies, the text contains 43 tokens (i.e., running words) and 31 types (i.e., different words), so

there were 1.39 tokens per type. Among the excerpt's words, most occur within the 1,000 most frequent word families in English. The words *listening, song, taught*, and *English* are from the list of the second most common 1,000 words in English, and only the words *study, twelve*, and *Australia* are outside these two frequency bands. The frequency for each word was also calculated. For example, the word *study* occurs 263 times per million words in the COCA corpus. These calculations, which were performed for each word in the excerpt, showed that on average, the excerpt's tokens occurred 724 times per million words, whereas the types occurred 840 times per million words. Each token in the excerpt had, on average, 1.3 syllables, an indication that the words that this particular speaker used were relatively short.

Based on the weighted clause ratio used to assess accuracy, the first sentence of the excerpt contains only a minor error (i.e., "start" instead of *started*). Because the AS-unit contains only minor errors that do not compromise meaning, it was coded as a "Level 1" error and given a score of 0.8. The AS-unit "I was taught her and" was coded as "Level 3" and was thus given a score of 0.1. The meaning is difficult to recover, since the utterance, when viewed in isolation, could mean *I taught her, I was teaching her*, or *I was taught by her*. As for fluency, the transcript represents the words produced in precisely 60 seconds, so the token count was used to calculate the number of words per minute. Among the tokens, 16 (37.2%) involved disfluencies, a fairly high rate.

As this example shows, the analysis was based on a wide range of measures. Some of these were objective. Others, such as the accuracy ratings, were inherently subjective.

3. Results

Of the 117 participants who completed interviews, nine were excluded from the study. Two participants in the personalized prompt (PP) condition and one in the impersonal prompt (IP) condition were excluded since they failed to respond to the prompt within a reasonable amount of time (30 seconds). Data from one participant in the IP condition was excluded, as he had grown up in the U.S. and therefore did not satisfy the recruitment criteria for the study. Another participant in the IP condition was excluded due to a technical failure that resulted in the Zoom session being cut short. Finally, four participants in the IP condition were excluded because they responded as if they were in the PP condition. This occurred in spite of the fact that both the English and Japanese directions explicitly stated that they were to talk about Japanese people in general and were to

avoid discussing their own experiences. It is noteworthy that none of the participants in the PP condition responded as if they had received the IP prompt.

3.1 Complexity

The transcripts of the remaining participants ($N = 108$) consisted of 57 in the PP condition and 51 in the IP condition. The first hypothesis (H1) predicted that participants in the IP condition would produce more complex output. In accordance with the hypothesis, the mean length of AS-units for the IP group ($M = 13.7$, $SD = 6.68$) was slightly higher than that of the PP group ($M = 13.1$, $SD = 6.01$). However, the result of an ANOVA with Prompt Type as between-subjects factor and mean length of AS-unit as the dependent variable did not show a significant difference between the groups, $F(1, 106) = 0.26$, $p = .610$, $\eta 2 < .01$. Thus, the results did not provide support for H1.

The descriptive statistics from other measures of complexity are shown in Table 2.

Table 2. Additional Complexity Measures for the Two Experimental Groups ($N = 108$)

	Personalized prompt (*SD*)	**Impersonal prompt (*SD*)**	***p***
Tokens per type	1.41 (0.27)	1.29 (0.27)	.002
Words within 1st 1000	58.4 (1.44)	59.9 (1.74)	.502
Frequency of tokens (*M*)	1131 (100)	1394 (138)	.120
Frequency of types (*M*)	1320 (123)	1488 (143)	.370
Syllables per word (*M*)	1.48 (0.02)	1.66 (0.04)	< .001

H1 predicts that participants in the IP condition would produce fewer tokens per type, a lower percentage of words within the 1st 1000 English words, words (both tokens and types) with lower average frequencies, and more syllables per word. Several ANOVAs were conducted to determine whether the differences between the two groups on these measures were significant. As predicted in H1, participants in the IP condition, compared to those in the PP condition, produced fewer tokens per type, $F(1, 106) = 10.23$, $p = .002$, $\eta 2 = .09$, while producing words with more syllables (a sign of lexical complexity), $F(1, 106) = 19.45$, $p < .001$, $\eta 2 = .16$. The other between-group differences related to words within the 1st 1000 words of English, $F(1, 106) = 0.45$, $p = .502$, $\eta 2 < .01$, the average frequency of each token, $F(1, 106) = 2.45$, $p = .120$, $\eta 2 = .02$, and the average frequency of each type, $F(1, 106) = 0.81$, $p = .370$, $\eta 2 = .01$, fell short of conventional significance at $p = .05$. Taken as a whole, these other measures provide only limited support for H1.

3.2 Accuracy

The weighted clause ratio (WCR) was used to assess the accuracy of each participant's ($N = 108$) production, with greater accuracy represented by numbers closer to 1.0. In line with H2, the PP group achieved slightly higher accuracy ($M = 0.511$, $SD = 0.242$) than the IP group ($M = 0.480$, $SD = 0.238$). However, an ANOVA, with group as a between-subjects factor and WCR scores as the dependent variable, indicated that this miniscule difference fell short of conventional significance, $F(1,106) = 0.46$, $p = .501$, $\eta 2 < .01$. Many studies of L2 production report errors in terms of the percentage of error-free AS-units. Using this measure, the results follow a similar pattern. The PP group had a mean of 17.1% ($SD = 25.5\%$) error-free AS-units, whereas the IP group had a mean of 14.08% ($SD = 23.0\%$) error-free AS-units. As with the WCR measure, this small difference fell short of conventional significance, $F(1,106) = 0.41$, $p = .521$, $\eta 2 < .01$. Thus, the results did not provide support for H2.

3.3 Fluency

The primary measure used to test H3 was the number of tokens produced per minute. To ensure that disfluencies such as false starts, repetitions, and self-corrections did not inflate this measure, tokens produced as part of a disfluency were omitted from the count. The PP group produced a mean of 37.4 tokens per minute ($SD = 18.0$), whereas the IP group produced a mean of 31.1 tokens per minute ($SD = 14.7$). An ANOVA with Prompt Type as the between-subjects factor and tokens per minute as the dependent variable indicated that the IP group's fluency, as assessed by this measure, was higher than that of the IP group, $F(1,106) = 3.97$, $p = .049$, $\eta 2 = .04$. As an additional measure of fluency, the percentage of tokens involved in disfluencies was calculated. For the PP group participants, a mean of 12.6% of the total tokens produced ($SD = 10.0\%$) occurred within disfluencies. For the IP group participants, a mean of 12.1% of the tokens ($SD = 12.8\%$) occurred within disfluencies. An ANOVA indicated that this miniscule difference was not significant, $F(1,106) < 0.01$, $p = .981$, $\eta 2 < .01$. The results thus provided support for H3 in terms of the number of tokens produced (excluding disfluencies).

3.4 Relationships between measures

To examine the relationships between the experimental measures, we carried out correlational tests between the measures for all participants ($N = 108$). To account for multiple comparisons in this part of the analysis, alpha value to reject the null

hypothesis was $a = .001$. As might be expected, the results indicated that greater complexity in terms of lexical sophistication was achieved at the expense of fluency. It can be seen taht, the number of syllables per word was negatively correlated with the number of disfluency-free tokens produced, $r(106) = -.428$, $p < .001$. Moreover, tokens per type (an indication of reuse of the same vocabulary items) had a strong positive correlation with the number of disfluency-free tokens produced, $r(106) = .662$, $p < .001$. However, syntactic complexity displayed the opposite pattern. It was found that, the mean length of AS-units was positively correlated with the number of disfluency-free tokens produced, $r(106) = .563$, $p < .001$. One possible interpretation of these findings is that the participants were relying on mostly automated knowledge instead of more controlled processes in their production of grammatical constructions but were relying on less automated knowledge when searching their memory for lexical items that conveyed their intended message. In terms of Levelt's model, the lexical and grammatical processing both occur within the Formulator. The results suggest that the two types of processing may be differentially affected by self-reference manipulations. The findings also suggest that controlled processing (Bialystok, 2001; McLaughlin, 1990) in the experimental task is focused primarily on lexical choices.

As would be expected, the weighted clause ratio (WCR) accuracy measure and the simpler measure in terms of error-free AS-units showed substantial overlap, $r(106) = .670$, $p < .001$. The WCR measure was negatively correlated with the percentage of disfluent tokens, $r(106) = .–318$, $p = .001$. This is to be expected since disfluency often occurred when participants corrected their own errors or repeated an ungrammatical utterance in order to hold the floor as they tried to come up with a more appropriate formulation of their message.

4. Discussion

The experimental results reported in this study suggest that personalized speaking prompts, relative to impersonal prompts, facilitate more fluent speech when fluency is operationalized as the number of words produced (with disfluent utterances omitted from the count). This finding is consistent with findings indicating that personal prompts similarly facilitate greater written production (Mueller, 2010, Exp. 2; Mueller & Kraus, 2018). However, the size of the effect in the current experiment was slightly smaller. In the Mueller and Kraus study, participants in the personalized prompt (PP) condition produced 26.6% more tokens, whereas the PP participants in the current study produced 20.3% more tokens.

These similar findings for L2 production, whether it involves speaking or writing skills, can be explained theoretically through reference to research on

memory and processing. Research on autobiographic memory (Conway, 2005) suggests that memories that are relevant to goals and self-defining experiences are highly accessible. In the experiment, participants were asked to describe their high school experiences. These memories would have been highly accessible due to both their recency and their relevance to current goals and personal interests. The memories were probably rendered even more salient as they were associated with the participants' experience of early adulthood, a time in which personal identity is forming. Research shows that memories from individuals' late-teens and early 20s, due to their relation to "pursuance of important goals that mark the emergence of an independent self-system in late adolescence", are highly accessible and remain so throughout the lifespan (Conway, 2005, p. 601). The memories' close association with identity also increases the likelihood that they are highly organized and elaborated, as suggested by research on the self-reference effect (Symons & Johnson, 1997).

Greater fluency in the personal prompt condition in the experiment can also be explained in terms of Levelt's (1989) model of speech production. Participants asked to talk about their high school experience could rapidly access a highly accessible and well-organized system of memories so as to rapidly form a message within the Conceptualizer. Participants asked to speak of Japanese people in general, on the other hand, would have presumably devoted more cognitive resources to deciding what to say (i.e., forming the message in the Conceptualizer). This would result in a lack of working memory resources available to assist in assembling the lexical and syntactic components in the Formulator. Participants' greater ease with personalized prompts was also suggested by patterns of failure to follow directions. Multiple participants in the impersonal prompt condition naturally switched to talking about themselves, even when told explicitly that this was not allowed. This suggests that some participants follow a strategy of switching the topic of an L2 conversation so that it focuses on personally relevant experiences and memories. The strategy may reflect participants' intuitions (which appear to be accurate) that it is easier to talk about personal experiences than it is to talk about impersonal facts.

The effects of prompt type on complexity and accuracy were less straightforward. Although the effects fell short of conventional significance, the trends in both the Mueller and Kraus (2018) study and the current study were roughly consistent in many respects with H1 (i.e., greater syntactic complexity and lexical sophistication in the impersonal prompt condition). This could be explained using Robinson's (2015) Cognition Hypothesis if it is assumed that impersonal tasks increase task complexity in a resource-directing manner. Within Robinson's taxonomy of task types, speaking tasks involving impersonal and personal speech prompts could be seen as varying on a perspective-taking dimension. According

to Robinson and Gilabert (2007) tasks differ depending on whether "the task requires the speaker/listener to take one first-person perspective on an event, or multiple second- and third-personal perspectives" (p. 165). It must be noted that this description of the perspective-taking dimension actually confounds several separate factors since reference (to self or others) and the number of perspectives adopted (as when the perspective shifts repeatedly) are separate task variables. The Mueller and Kraus (2018) study found that personalized prompts were associated with fewer errors. While the current study found a slight trend in this direction, the error rates were essentially the same for both conditions.

One of the more interesting findings in the current study concerned the relationship between fluency and complexity, regardless of the prompt type. Whereas higher syntactic complexity was associated with greater fluency, higher lexical complexity was associated with reduced fluency. The finding may reflect learners' default strategies when speaking. Among the four skills, speaking is particularly demanding due to time pressure. The speaker must formulate utterances rapidly before the listener loses interest in the conversation or loses the thread of the discussion. To deal with time pressures, L2 speakers may avoid trying out grammatical patterns that are not highly automatized. This may result in speech that does not take full advantage of the range of syntactic patterns available to express an idea. This speaking strategy, which seeks to avoid unwarranted risks, may emerge from situational constraints associated with spoken interaction. After all, many ideas can be expressed with fairly basic syntax. When selecting lexical resources, on the other hand, speakers have fewer options. Some ideas can be expressed with multiple words and expressions, but quite often, there are few lexical alternatives, and this applies *a fortiori* to L2 learners who have much less lexical knowledge than native speakers. In such cases, the learner is likely to expend critical cognitive resources searching for vocabulary that can suitably convey their message. Behaviorally, this could result in their speech slowing down or halting as they consider various potential lexical items.

5. Conclusion

Speaking prompts play a major role in determining whether a speaking activity is motivating and is suited to a learner's current level of competence. For this reason, more research is needed to determine how prompt variables affect learners' L2 acquisition, L2 performance, and subjective impression of a task. The fact that some students resorted to self-reference even when instructions explicitly stated that self-reference should not be used suggests an L2 speaking strategy worthy of further exploration. This paper has focused on the ease with which students

access language related to self-reference, but another avenue to investigate would be the potential role of self-reference prompts in increasing learners' motivation during speaking tasks.

Future research should also consider several possible improvements over the current study's design. First, this study used a between-subjects design. The study's relatively high *N*-size made this feasible; however, differences in participants' speaking ability, apparent in the results, led to high variance across the measures, which in turn, resulted in the experimental measures having less than optimal sensitivity. To minimalize the effect of variance in proficiency across participants, future researchers may want to use within-subjects designs.

An additional issue concerns measures such as the type-token ratio. This measure is highly sensitive to the number of tokens produced, and as a result, participants who produce fewer words have artificially high type/token ratios. The issue can be partially ameliorated through the use of D, an alternative lexical diversity measure (Malvern & Richards, 1997), but this option is not available for shorter texts such as the transcripts analyzed in the current experiment. Future researchers may therefore want to consider reducing the *N*-size while increasing the time on task so that participants produce more tokens on average.

At a more general level, future research on the self-reference effect on L2 output, to include both writing and speaking, should examine the effects of moderator variables associated with participants' age as well as their prior age during the recalled events. Previous research (Jansari & Parkin, 1996; Munawar et al., 2018) has shown that there is an episodic reminiscence bump in which episodic self-knowledge from one's teens and early 20s is more accessible than self-knowledge from other periods of life. The effect is attributed to stronger encoding of events in adolescence, which results in more frequent retrieval of the stored memories, with frequent retrieval leading to further strengthening of the memories (Janssen et al., 2005). The relevance of the reminiscence bump to second language production has not been investigated yet, but one hypothesis would be that output related to events and experiences occurring in adolescence (as in the self-reference conditions in the current study and the Mueller & Kraus, 2018, study) would be subject to stronger self-reference effects.

The current study found that L2 learners speak more when given personalized prompts. This suggests that prompts should be personalized when L2 learners exhibit hesitancy to speak or actively engage with classroom tasks. The findings indicated that learners' search for the appropriate words to express a message often reduced their fluency. This would suggest that instructors should use personal prompts while manipulating other factors to reduce task complexity when the emphasis is on learners' development of fluency through the automatization of existing lexical and syntactic knowledge. Of course, impersonal tasks also have

their place since L2 learners must have the ability to discuss topics unrelated to their own personal experiences. These tasks can be rendered less challenging through the use of task progressions that move from simpler to more complex tasks. For example, impersonal speaking tasks may be preceded by related personalized tasks, or by listening tasks in which essential vocabulary and grammatical patterns are introduced.

References

Bellezza, F.S. (1984). The self as a mnemonic device: The role of internal cues. *Journal of Personality and Social Psychology*, *47*(3), 506–516. https://doi.org/10.1037/0022-3514.47.3.506

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511605963

Bui, G., & Skehan, P. (2018). Complexity, accuracy, and fluency. In J.I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). John Wiley and Sons. https://doi.org/10.1002/9781118784235.eelt0046

Bygate, M. (2001). Effects of task repetition on the structure and control of oral interaction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogical tasks, second language learning, teaching and testing* (pp. 23–48). Longman.

Chang, M.-M., & Lehman, J.D. (2002). Learning foreign language through an interactive multimedia program: An experimental study on the effects of the relevance component of the ARCS model. *CALICO Journal*, *20*(1), 81–98. https://doi.org/10.1558/cj.v20i1.81-98

Conway, M.A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. https://doi.org/10.1016/j.jml.2005.08.005

Conway, M.A., & Loveday, C. (2015). Remembering, imagining, false memories. *Consciousness and Cognition*, *33*, 574–581. https://doi.org/10.1016/j.concog.2014.12.002

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University.

Ferguson, T.J., Rule, B.G., & Carlson, D. (1983). Memory for personally relevant information. *Journal of Personality and Social Psychology*, *44*(2), 251–261. https://doi.org/10.1037/0022-3514.44.2.251

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

Gaies, S.J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly*, *14*(1), 53–60. https://doi.org/10.2307/3586808

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins. https://doi.org/10.1075/lllt.32.01hou

Hunt, K.W. (1970). Recent measures in syntactic development. In M. Lester (Ed.), *Readings in applied transformational grammar*. Holt, Rinehard and Winston.

Jansari, A., & Parkin, A.J. (1996). Things that go bump in your life: Explaining the reminiscence bump in autobiographical memory. *Psychology and Aging*, *11*(1), 85–91. https://doi.org/10.1037/0882-7974.11.1.85

Janssen, S.M.J., Chessa, A.G., & Murre, J.M.J. (2005). The reminiscence bump in autobiographical memory: Effects of age, gender, education, and culture. *Memory*, *13*(6), 658–668. https://doi.org/10.1080/09658210444000322

Keller, J.M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, *10*(3), 2–10. https://doi.org/10.1007/BF02905780

Kerz, E., & Wiechmann, D. (2020). The role of verbal working memory in L2 sentence comprehension. *Journal of Second Language Studies*, *3*(1): 1–30. https://doi.org/10.1075/jsls.18022.ker

Kihlstrom, J.F., Cantor, N., Albright, J.S., Chew, B.R., Klein, S.B., & Niedenthal, P.M. (1988). Information processing and the study of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 21* (pp. 145–180). Academic Press. https://doi.org/10.1016/S0065-2601(08)60226-9

Klein, S.B. (2012). The self and its brain. *Social Cognition*, *30*(4) 474–518. https://doi.org/10.1521/soco.2012.30.4.474

Klein, S.B., & Kihlstrom, J.F. (1986). Elaboration, organization, and the self-reference effect in memory. *Journal of Experimental Psychology: General*, *115*(1), 26–38. https://doi.org/10.1037/0096-3445.115.1.26

Klein, S.B., & Loftus, J. (1988). The nature of self-referent encoding: The contributions of elaborative and organizational processes. *Journal of Personality and Social Psychology*, *55*(1), 5–11. https://doi.org/10.1037/0022-3514.55.1.5

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. MIT.

Malvern, D.D., & Richards, B.J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language: papers from the annual meeting of the British Association for Applied Linguistics held at the University of Wales, Swansea, September 1996* (pp. 58–71). Multilingual Matters.

Mayer, R.E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology*, *96*(2), 389–395. https://doi.org/10.1037/0022-0663.96.2.389

McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11*(2), 113–128. https://doi.org/10.1093/applin/11.2.113

Michel, M.C. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Routledge. https://doi.org/10.4324/9781315676968-4

Moreno, R., & Mayer, R.E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology*, *92*(4), 724–733. https://doi.org/10.1037/0022-0663.92.4.724

Mueller, C.M. (2010). Effects of explicit instruction on incidental noticing of metaphorical word sequences during a subsequent reading task. *International Journal of English Studies*, *10*(1), 81–101. https://doi.org/10.6018/ijes/2010/1/113991

Mueller, C.M., & Kraus, W.A. (2018). The effects of personalized prompts on Japanese EFL students' written essays. *OnCUE Journal*, *11*(1), 25–50.

Munawar, K., Kuhn, S. K., & Haque, S. (2018). Understanding the reminiscence bump: A systematic review. *PLoS ONE*, *13*(12), e0208595. https://doi.org/10.1371/journal.pone.0208595

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), 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogical task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT. A decade of plenaries from the international conference* (pp. 87–121). John Benjamins. https://doi.org/10.1075/tblt.8.04rob

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 161–176. https://doi.org/10.1515/iral.2007.007

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*(9), 677–688. https://doi.org/10.1037/0022-3514.35.9.677

Skehan, P. (1998). *A cognitive approach to second language learning.* Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185–211. https://doi.org/10.1177/136216889700100302

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive process they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391. https://doi.org/10.1093/applin/16.3.371

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Text Inspector. (2018). Online lexis analysis tool at textinspector.com

Thomsen, D. K. (2015). Autobiographical periods: A review and central components of a theory. *Review of General Psychology*, *19*(3), 294–310. https://doi.org/10.1037/gpr0000043

Wright, C. (2020). Effects of task type on L2 Mandarin fluency development. *Journal of Second Language Studies*, *3*(2): 157–179. https://doi.org/10.1075/jsls.00010.wri

Wu, S., & Ma, Z. (2016). Suppression and working memory in auditory comprehension of L2 narratives: Evidence from cross-modal priming. *Journal of psycholinguistic research*, *45* (5):1115–1135. https://doi.org/10.1007/s10936-015-9390-2

Address for correspondence

Charles M. Mueller
Fuji Women's University
Toyohira-ku Nakanoshima 2jo 1-1-17-409
062-0922 Hokkaido Sapporo-shi
Japan
mueller@fujijoshi.ac.jp
https://orcid.org/0000-0002-3369-9124

Co-author information

Peter Richardson
Hokkaido Bunkyo University
peterrichardson@do-bunkyodai.ac.jp
https://orcid.org/0000-0003-0621-7879

Publication history

Date received: 11 June 2021
Date accepted: 21 November 2021
Published online: 31 December 2021