The Effects of Personalized Prompts on Japanese EFL Students' Written Essays

Charles M. Mueller and William A. Kraus

Fuji Women's University

Personalized writing prompts are often used in second language pedagogy to foster more meaningful learning. Unfortunately, few studies have examined the effects such prompts have on learners' written output. The current research discusses two research strands that may have relevance when considering the effects of such prompts on written products: memory research on the self-reference effect and Robinson's Cognition Hypothesis. To empirically investigate the effects of writing prompt manipulations on output, an experiment was conducted with 36 Japanese-L1 female EFL participants whose written essays were examined in terms of verbosity, complexity, and accuracy. Based on memory research and the Cognition Hypothesis, it was hypothesized that personalized writing prompts would lead to greater verbosity, less complexity, and less accuracy. While the findings were partly consistent with these hypotheses, it was found that personalized prompts were, in fact, associated with greater accuracy. The results suggest that personalized prompts may be a preferable pedagogic option when the focus is on promoting written fluency. Prompts asking learners to write about other people or to take an impersonal stance, on the other hand, may require greater scaffolding and may need to be introduced later in the task cycle.

第二言語教授法において，ライティングで個人的な内容を指示文(Prompts)とすることは，より意義深い学習を促すためによく利用されている．ところが残念なことに，このような指示文が、学習者の書いた成果物にどのような効果をもたらすかはあまり調査されていない．本研究では，その効果に関連しうる二つの研究－記憶の自己参照効果，およびロビンソンの認知仮説を考拠とした．ライティングにおける個人的な内容の指示文の効果を実験的に調査するため，英語を外国語として学習する女性日本語母語話者36 名の参加者によって書かれたエッセイを用い，文章の長さ，複雑さ及び正確さを分析した．記憶研究および認知仮説に基づき，個人的なことを書く場合はより長い文章になる一方，複雑さや正確さはより劣るだろうという仮説が立てられた．結果は仮説と一致する部分が多かったものの，正確さについてはより

高まることがわかった. したがって, ライティングの流暢さに焦点を当てる場合は, 個人的な内容を指示文とする方がより好ましい教授的選択であると言える. 他方, 他者に関することや非個人的な立場をとる内容を指示文とする際は, より強力な足場づくりが必要となり, 学習が進んでから用いるのが望ましいと考えられる.

The ability to write in English is a concern for nonnative speakers (NNS) of English in a world that increasingly uses English in a wide range of situations, such as international business interactions and communication on social networks. L2 writing is particularly relevant within the Japanese English as a Foreign Language (EFL) context due to reports that Japanese learners receive inadequate training in L2 English writing practices in university (Fujii, 2012) and often fail to improve their writing ability. Sasaki (2011), for example, examined Japanese college students' English writing development over a three-and-a-half-year period. She found that the composition scores for the fourth-year students who did not do study abroad were, on average, virtually identical to their composition scores during their first year.

To address this need for better L2 writing instruction, teachers have sought to develop more effective pedagogical tasks. In the case of writing, these generally involve a writing prompt. Teachers have often preferred personalized prompts that invite learners to write and talk about themselves. This is understandable. After all, in the general field of education, personalized activities have often been justified on both affective and cognitive grounds. Use of the self as a reference point is said to increase engagement and encourage learners to "use available cognitive capacity for active cognitive processing of the incoming information during learning," leading to deeper processing that "results in more meaningful learning as indicated by better transfer on test performance" (Mayer, Fennell, Farmer, & Campbell, 2004, p. 391; see also Moreno & Mayer, 2000). In L2 instructional contexts, this assumption drives many pedagogical choices, such as the extensive use of personalized tasks within mainstream foreign language textbooks.

Personalized tasks also appear in pedagogical sequences with explicit psycholinguistic justification. To cite just one example, VanPatten's (2002) processing instruction cycle of pedagogical tasks involves (1) explicit instruction

regarding learners' faulty understanding of a target form, (2) the use of an input task in which correct comprehension of the target form is essential to complete the task, and then (3) the use of output (typically speaking) activities in which learners produce the target form within the context of a personalized task.

In second language acquisition (SLA) research, personalized tasks can perhaps be justified on general assumptions, namely, the idea that education should promote activities and content that are personally relevant to learners (for a general theoretical framework emphasizing personal relevance, see Keller, 1987; for an empirical study examining the relationship between task relevance and learning in L2A, see Chang & Lehman, 2002). However, personalized tasks should ideally also receive justification in terms of empirically assessed outcomes.

This paper thus examines the effects of using a personalized prompt in an essay writing task. More specifically, it reports an experiment comparing written outcomes when participants respond to a self-reference (SR) prompt versus an other-reference (OR) prompt. SR prompts will be defined as prompts eliciting a first-person vantage point (e.g., "What do you like to do on the weekend?"), and OR prompts as those eliciting reference to the actions and thoughts of another person or an impersonal vantage point (e.g., "What do Japanese people like to do on the weekend?").

The manipulation of writing prompts along a personalized/depersonalized dimension is predicted to have effects on the verbosity, complexity, and accuracy of L2 writers' output. These predictions are based on four strands of cognitively oriented research focused on (1) cognitive writing processes, (2) the self and memory, (3) autobiographical memory, and (4) task difficulty.

Cognitive writing processes have been described in Flower and Hayes' (1981) influential model, which divides the writing process into "planning", "translating", and "reviewing" sub-processes (p. 370). *Planning* consists of generating ideas by retrieving relevant information from long-term memory and then organizing this information while attending to writing goals. *Translating* is the process of transforming meaning into a linear piece of writing. *Reviewing* involves examining one's written text in order to make revisions or as a "springboard to further translating" (p. 374).

In this paper, it is assumed that personalized tasks facilitate the writing process through the enhancement of the aforementioned planning and translating sub-processes. SR may ease learners' cognitive load by reducing the need for planning so that more attention can be given to composing language (i.e., in spoken production, this would involve processing within the formulator, see Levelt, 1989; see also Skehan, 2009). The self is said to provide an extensive framework of internal cues organized in the form of rich, coherent, and highly accessible autobiographical memories (Conway, 2005), which play a crucial role in the construction of personal identity and facilitate problem-solving and the maintenance of close interpersonal relationships (Harris, Rasmussen, & Berntsen, 2013). Research has shown that material associated with the self is more readily retrieved during recognition and recall tasks (Bellezza, 1984; Conway & Holmes, 2004; Symons & Johnson, 1997). If this is the case, personalized prompts should reduce the cognitive load associated with planning processes when students are engaged in L2 writing tasks.

The translating process may also be facilitated by personalized tasks. This process, which involves employing linguistic knowledge from long-term memory to express conceptual content, can be especially arduous for L2 writers whose language processing is often slow (for a discussion of automatization and SLA, see Segalowitz, 2003). As Schoonen and his colleagues (2003) point out, slowed access to lexis and grammatical structures "will burden the working memory and thus hinder the writing process as such, not just with respect to writing fluency, but also with consequences for the quality of the text" with detrimental effects "larger for L2 writing than for L1 writing" (p. 171).

Perhaps aware of the facilitative role of personalized tasks, EFL materials used throughout the world often have students write or talk about themselves. To cite just one example, Richards (2006), in a justification of the design features of an English textbook for secondary students, states as one of the "course principles" that the textbook provides "personalization tasks" that "offer opportunities for students to use the language to speak about themselves" (p. 15). Since typical L2 writers in junior high, high school, and college are often asked to write and speak about themselves in the L2, the linguistic resources and routines associated

with the self are likely to be more practiced and automatic. As Segalowitz (2003) notes, increased automaticity is, in turn, associated with reduced demands on cognitive resources, more fluent and seamless use of language, and the ability to employ linguistic resources without conscious attention.

Predictions regarding the personalization of L2 writing prompts are also informed by work on task difficulty and task complexity. One of the most influential taxonomies of task characteristics in SLA has been put forward as part of the Cognition Hypothesis, which maintains that SLA is facilitated by a systematic progression that moves from simple to more complex tasks (Robinson, 2001, 2015). This framework is based on earlier theoretical work (e.g., Slobin, 1993) that has examined task complexity in L1 acquisition. Within Robinson's framework, increased task complexity can be either "resource directing" or "resource dispersing." Increases in the resource-directing dimension of tasks are thought to promote noticing (Leow, 2015; Schmidt, 2001) and internalization of the language forms employed to meet increased task demands. An example of the resource-dispersing dimension would be the presence or absence of time to plan what to say. Greater task complexity along resource dispersing dimensions is said to be helpful in promoting greater automatization of a learner's current linguistic resources.

Robinson (2005) has argued that tasks that are difficult along resource-directing dimensions "lead the learner to attempt to map the increasing conceptual/functional requirements of tasks onto speech, in such a way as to affect fluency negatively, but, in selected domains, to facilitate the development of increased accuracy and complexity in production" (p. 7). Skehan (1998, 2009), on the other hand, claims that increased task demands degrade fluency, complexity, *and* accuracy. According to his Trade-Off Hypothesis, limitations in working memory resources force learners to differentially allocate resources so as to satisfy competing demands of each given task. It should be noted that the term *fluency*, as employed by these and other SLA researchers, refers to a "multidimensional construct" related to the speed of production, the number of breakdowns (e.g., pauses), and the number of repairs, such as false starts (Housen, Kuiken, & Vedder, 2012, p. 5).

Robinson (2007) broadly divides pedagogical L2 tasks into cognitive, interactive, and learner factors. Within this taxonomy, self-reference (SR), while not mentioned explicitly as a factor, would be closely associated with "perspective-taking," a cognitive factor described as resource-directing. According to Robinson and Gilabert (2007), tasks vary depending on "whether the task requires the speaker/listener to take one first-person perspective on an event, or multiple second- and third-person perspectives" (p. 165). It should be noted that their perspective-taking factor combines what are, in fact, several separate factors that depend on (1) reference (the self, another familiar person, etc.), (2) the number of perspectives taken, as when conversations shift repeatedly from one perspective to another, and (3) whether one is speaking of oneself or listening to others speak of themselves.

Based on Robinson's work, it will be assumed in this paper that the SR/OR manipulation of a writing prompt differentially affects task complexity. Productive L2 tasks that involve SR as opposed to OR should result in greater fluency, whereas OR should be more cognitively demanding and should therefore result in less fluency but more complex language. Robinson (2005) claims that tasks that increase resource-directing dimensions "draw learner attention to the ways in which the L1 and the L2 may differentially *grammaticize* conceptual notions" and thus "have positive effects on L2 accuracy of production" (p. 9, italics in the original). This paper will therefore assume that OR prompts are associated with fewer errors.

Considerable research has been conducted on the effects of task characteristics on L2 writing (Kuiken & Vedder, 2007, 2008, 2011; Ong & Zhang, 2010; Révész, Kourtali, & Mazgutova, 2016), yet there appears to be little or no research that has specifically explored personalized prompts in terms of their effects on L2 writing. More research is thus needed to determine the effects of personalization on outcomes related to L2 input and output activities. The current research examines personalization effects on L2 output, focusing on essay writing, as this is a common output task in college-level EFL classes in Japan.

Only L2 output is examined in the current study. This type of measure does

not provide fine-grained insights into participants' productive processes such as revision. For this reason, participants' writing fluency will be examined indirectly through measures of verbosity (i.e., how much language participants produce in their written text). As an indirect measure of writing speed, verbosity can be regarded as an imperfect proxy for fluency.

This study specifically examines the effects of personalization to address three research questions:

RQ 1: Are personalized writing prompts associated with more verbose responses?

RQ 2: Are impersonal writing prompts associated with more complex output?

RQ 3: Are impersonal writing prompts associated with greater accuracy?

Method

Participants

The participants were 36 Japanese-L1 female college freshmen in a required first-year writing class taken during their first semester. They were all in a department that focused on English linguistics and literature. Based on the TOEFL PBT scores of the college's incoming class, most were probably at the B1 level in terms of the Common European Framework of Reference (CEFR). They were motivated learners, many of whom aspire to enter a career requiring English such as work as a flight attendant or as an English teacher. Participants from two classes, both taught by native-English speakers (the authors), were randomly assigned to the experimental blocks.

Procedure

The participants wrote about two topics (health and high school life) in either a "self-reference" condition or "other-reference" (OR) condition in response to two of the four following prompts:

A. [SR: health prompt] Describe what you do to stay healthy. (You should talk only about yourself. Do NOT describe the experiences of other family members or Japanese people in general.)

B. [OR: health prompt] Describe what Japanese people typically do to stay healthy. (You should talk about typical Japanese people. Do NOT talk about yourself.)

C. [SR: high school prompt] Describe your high school experiences. (You should talk only about yourself. Do NOT describe the experiences of other family members or Japanese people in general.)

D. [OR: high school prompt] Describe a typical Japanese person's high school experiences. (You should talk about typical Japanese people. Do NOT talk about yourself.)

Table 1 shows how the four blocks of participants were set up so that all received only one prompt in the SR condition and only one prompt in the OR condition. Moreover, each of the four prompts appeared an equal number of times resulting in nine students per block. The sequencing of the topics was also counter-balanced so that half of the participants did the SR condition first and half did it last.

The participants did all the writing individually in the university's computer lab. They were not allowed to use dictionaries or to reference the Internet while writing. The QUIA (https://www.quia.com/) online testing platform was used to provide the participants with their prompts and to record responses. The participants were told that they would have 20 minutes to respond to each prompt. An on-screen countdown clock within the QUIA site helped them keep track of remaining time. After 20 minutes, the QUIA site automatically saved

Table 1

Counter-Balancing of Conditions in Data Collection Procedure

	First Prompt	Second Prompt
Block A	SR[a] (health)	OR[b] (high school)
Block B	SR (high school)	OR (health)
Block C	OR (high school)	SR (health)
Block D	OR (health)	SR (high school)

[a] SR = self-reference

[b] OR = other-reference

the participants' responses and stopped recording any additional input from the computer keyboard, so the participants were unable to exceed the allotted time.

Analysis

Prior to data analysis, the spelling in the participants' essays was corrected. No instances in which the participants' misspelling resulted in ambiguity were noted. Even if the participant's original word was an actual English word (e.g., *lice*), this was corrected if the intended word was clearly something else (e.g., *rice*). The spell-corrected essays were used for all subsequent analyses. The participants' essays were analyzed in terms of various measures related to verbosity, complexity, and accuracy. The essays were initially analyzed using an online vocabulary profiler (Web Vocabprofile at http://www.lextutor.ca/vp/eng/ by Cobb; based on the Range program by Heatley & Nation, 2002) and another online site (https://readability-score.com/text/) for readability scores and related measures.

Verbosity. Within the context of the current study, fluency will be operationalized in terms of verbosity (i.e., tokens produced within a set amount of time). Previous SLA research (e.g., Way, Joiner, Seaman, 2000) has generally employed word counts as measures of fluency; yet it should be noted that word counts are an indirect measure of fluency. Unlike spoken fluency measures, which include fine-grained features such as false starts, word count measures focus exclusively on the written product and do not directly assess writers' behavior as they seek to employ language rapidly and smoothly. The term *verbosity* has therefore been used in this paper to indicate that the measure of participants' fluency is indirect.

To determine verbosity so as to answer the first research question, each essay was analyzed in terms of the total number of tokens. Additional analyses tallied the total number of (1) types (i.e., distinct words in the text), (2) characters, (3) syllables, and (4) sentences. High scores on these measures, for the most part, reflect a participant's ability to provide a long response in the short 20 minutes of allotted time. In other words, the measures are likely to reflect writing fluency, suggesting greater ease in responding to the prompt. A high number of types, in addition to reflecting verbosity, signifies less repetition.

Complexity. To measure complexity and answer the second research question, the mean length of T-units was calculated manually by both authors. The T-unit measure was originally created by Hunt (1965) as an index of syntactical sophistication in children's L1 writing. It is defined as "a main clause plus all subordinate clauses and nonclausal structures attached to or embedded in it" (Hunt, 1970, p. 4). One reason the T-unit was originally developed was to counteract the failure of raw sentence counts to take into account the effects of run-on sentences. By breaking such sentences into separate units, the T-unit, Hunt reasoned, was a better basic measure of syntactic complexity.

Research suggests that the length of T-unit shows a positive correlation with writing proficiency (Ortega, 2003). As would be expected, the mean length of T-units (MLTU) correlates highly with the mean length of a sentence ($r = .907$, in Lu, 2010). In the current study, the T-unit determinations by the authors were subsequently checked using a program (http://aihaiyang.com/software/l2sca/batch/) created by Ai and Lu (2013). However, the automated program was found to be of little use as it often made obvious mistakes when analyzing sentences containing errors. (This is understandable since the program was not originally designed for the analysis of non-native texts.)

It should be noted that some researchers (e.g., Gaies, 1980) have questioned the appropriateness of using the T-unit with low-proficiency learners due to difficulties in determining T-units in texts with numerous errors. In the current study, it was decided to use the T-unit in spite of such concerns based on the following considerations: (1) use of the T-unit makes the findings easier to compare with other research and (2) both researchers' subjective impressions were that the participants' written texts could be divided into T-units with little ambiguity. That is to say, the errors, while numerous, virtually never resulted in instances in which the participant's intended meaning could not be ascertained with a high degree of confidence.

Subsequently, additional measures of complexity and lexical use were examined. Each essay was examined in terms of:

(1) lexical density (i.e., the percentage of content words)
(2) tokens per type

(3) tokens per word family

(4) characters per word

(5) syllables per word

(6) words per sentence

(7) Flesch-Kincaid Grade Level

(8) Coleman-Liau Index (Coleman & Liau, 1975)

(9) New Dale-Chall Formula (Chall & Dale, 1995),

(10) vocabulary profile (analyzed using the Vocab Profiler at http://www. lextutor.ca/vp/eng/).

The Flesch-Kinkaid Grade Level index is a composite measure based on the average number of syllables per word and the average sentence length in a text (Kincaid, Fishburne, Rogers, & Chissom, 1975). The Coleman-Liau Index and New Dale-Chall Formula are composite measures of readability based on a slightly different (and overlapping) set of factors. Since most of these measures were not independent or (as in the case of the vocabulary profile) were coarse, the analyses related to these measures were of an exploratory nature.

Accuracy. The essays were also coded for level of accuracy. In this paper, accuracy refers to the appropriate use of morphosyntax and lexis to convey meaning. Researchers have struggled to develop useful and valid measures of accuracy. One issue with using raw tallies of errors or counts of errors per linguistic unit (e.g., per T-unit) is that the tallies do not reflect the relative seriousness of the error (Polio, 1997). Classification of errors, while time-consuming, would appear to be ideal; however, there are practical problems when analyzing texts by low-proficiency writers, as the writer's intent (and thus the source of the error) is not always clear. If we read the sentence *I saw a dogs*, for example, it is not clear whether the indefinite article or the plural marking on *dogs* is superfluous. For this reason, in the current study, accuracy was measured using a modified version of the weighted clause ratio created by Foster and Wigglesworth (2016), a measure that assesses the level of accuracy of each clause. These researchers used a scale ranging from 0 (no errors) to 3. When clauses contained multiple errors, they coded the clause based on the most serious error. Because the essays in the current study contained a high number of errors, it was felt that a more coarse-

grained analysis would be more practical and accurate. We therefore adapted the Foster and Wigglesworth formula *to T-units* as follows:

0 = The T-unit has no errors at all.

1 = The T-unit has only minor errors that do not interfere at all with determining the meaning.

2 = The T-unit has serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always immediately obvious.

3 = The T-unit has very serious errors that make the intended meaning far from obvious and only partly recoverable.

It should be noted that the current researchers' knowledge of Japanese along with their awareness of SLA processes and typical Japanese patterns of error enabled them to recover the meaning even when the error was quite egregious. For example, it would be clear that if a participant mentioned moving to another "mansion," she had actually intended to use the word "apartment" but had inadvertently transferred the meaning of the Japanese false friend *manshon* (apartment). A native speaker unfamiliar with Japanese learners' typical patterns of error due to crosslinguistic transfer would completely misunderstand the intended meaning in this case and the intended meaning (i.e., apartment) would not be recoverable, making this a Level 3 error. The two authors, on the other hand, would, in this case, realize the participant's intended meanings. For this reason, to ensure that the errors were assessed objectively in a manner that would be replicable across studies, the distinction between Level 2 and Level 3 errors was made based on the comprehensibility of the T-unit to an idealized naïve English reader (i.e., a native speaker without extensive experience working with Japanese students).

Even with these caveats, it must be noted that the analysis of the participants' errors was often subjective, as considerable inferencing was often necessary to determine the intended meaning. To offer just one example, one participant, after stating the importance of taking baths for health, concluded her essay with the statement, "Japan is famous for spring." While the sentence would appear to be free of errors when viewed in isolation, context would suggest that the participant was actually referring to hot springs. For this reason, the ratings

were all performed by both authors. There was a strong correlation between the two sets of error ratings, $r = .88$, $p < .001$. The reported results are based on the averages of both authors' ratings for each T-unit.

Results

The results of the comparison of the 36 participants' two essays belonging either to the Self-Reference (SR) condition or the Other-Reference (OR) Condition are here discussed under three headings based on whether the measures are related primarily to verbosity, complexity, or accuracy.

Verbosity

As mentioned previously, verbosity can be viewed as an indirect measure of a writer's fluency. To determine whether the SR condition was associated with greater verbosity, counts were made of tokens (i.e., running words) in essays in both conditions. As can be seen in Table 2, participants in the SR condition consistently produced longer texts. Notably, in terms of the total word count (i.e., tokens), the SR texts were about a quarter (26.6%) longer. A paired samples *t*-test (Table 2) showed this difference to be significant ($p = .001$) at an alpha value level of .017 (i.e., with alpha subjected to a Bonferroni adjustment, taking into account the three key experimental measures related to verbosity, complexity, and accuracy which were used to answer the three research questions put forth in the introduction). The answer to the first research question then is affirmative: personalized writing prompts are associated with more verbose responses.

Further analyses were conducted to determine whether greater verbosity was evident using other related measures (i.e., token, type, character, syllable, and sentence counts). Paired-sample *t*-tests were also conducted for these measures. On all these measures, except for total characters ($p = .051$), there was a significant difference between the SR and OR condition at $p = .05$, with greater verbosity seen in the SR condition. It should be noted that these measures are not independent; hence they cannot be used to directly address the first research question.

Table 2

Verbosity Measures for the Self-Reference and Other-Reference Conditions

	SR [a]	OR [b]		95% CI of	Significance	Cohen's
	M (SD)	*M* (SD)	*t* (df)	Difference	(2-tailed)	*d*
Tokens	110.4 (43.7)	87.2 (29.9)	3.62(35)	10.2 to 36.2	*p* = .001	0.603
Types	65.1 (18.8)	54.2 (14.8)	4.34(35)	5.8 to 16.0	*p* < .001	0.723
Characters	447.7 (175.5)	395.9 (137.9)	2.02(35)	-0.2 to 103.8	*p* = .051	0.337
Syllables	136.0 (54.7)	118.4 (40.2)	2.28(35)	1.9 to 33.1	*p* = .029	0.380
Sentences	10.0 (3.8)	7.7 (2.2)	4.15(35)	1.1 to 3.4	*p* < .001	0.692

[a] SR = self-reference

[b] OR = other-reference

Complexity

As can be seen by the MLTU measure in Table 3, the OR condition was associated with longer T-units (9.7 tokens). However, the difference between the SR and OR conditions was just short of significance at an alpha value level of .017. The experiment thus failed to provide clear evidence that impersonal prompts are associated with more complex output. However, it should be noted that failure to show significant results, in this case, reflects the loss of sensitivity (i.e., the use of an alpha of .017 to control for experiment-wise error) that is inevitable in the current experimental design, which included multiple measures within the same experiment. To determine whether other measures would indicate a difference in complexity between the two conditions, a mean length of sentence count was performed on the SR and OR texts. In this case, the differences were not significant (Table 3).

In further analyses, other complexity measures were calculated. Complexity is associated with higher use of content words relative to function words, fewer tokens per type or per word family (i.e., greater lexical diversity), more characters and syllables per word (i.e., use of longer words), and more words per sentence (i.e., longer sentences). More complex writing employs more low-frequency lexical items, resulting in less use of the 1000 most common word families in

Table 3

Syntactic Complexity Measures for the Self-Reference and Other-Reference Conditions

Measure	SR [a] M (SD)	OR [b] M (SD)	t(df)	95% CI of Difference	Significance (2-tailed)	Cohen's d
T-units	12.4 (4.3)	9.1 (2.7)	5.03(35)	2.0 to 4.6	p < .001	0.838
MLTU [c]	9.0 (1.6)	9.7 (1.8)	-2.43(35)	-1.3 to -0.1	p = .020	0.404
Sentences	10.0 (3.8)	7.7 (2.2)	4.15(35)	1.1 to 3.4	p < .001	0.690
MLS [d]	11.5 (3.5)	11.9 (5.9)	-0.36(35)	-2.3 to 1.6	p = .720	0.060

[a] SR = self-reference

[b] OR = other-reference

[c] Mean length of T-unit

[d] Mean length of sentence

English and more use of "off-list" words (lexical items that are not in the first two-thousand word families of English or the Academic Word List) and in many cases, more use of academic words. As can be seen in Table 4, on measures showing a significant difference between the two conditions, the OR condition is associated with greater lexical complexity. The texts in this condition had significantly more content words (i.e., greater lexical density), fewer tokens per word family (a sign of greater lexical diversity), and more characters and syllables per word (indicating longer words). Both the Flesch-Kincaid Grade Level and New Dale-Chall Formula showed a significant difference between the two conditions. The former places the OR texts around the sixth-grade level and the SR texts between the fourth- and fifth-grade levels. The Coleman-Liau Index gives a higher estimate for the grade level of both sets of texts, but once more, with a significantly higher grade level for the OR texts.

As shown in Table 4, the OR texts had similar tokens per type compared to SR texts but fewer tokens per family. Lexical density is characteristic of academic prose, but it can also be the result of some common patterns of omission among L2 users, as when learners omit English articles and prepositions. Fewer tokens per type and fewer tokens per family are features generally associated with less repetition and less anaphoric use of synonyms. In the results, the OR texts had

Table 4

Lexical Complexity Measures for the Self-Reference and Other-Reference Conditions

Measure	SR[a] M (SD)	OR[b] M (SD)	t(df)	95% CI of Difference	Significance (2-tailed)
Content words	48.9% (0.04%)	57.0% (0.07%)	-5.42(35)	-11.1% to -5.1%	p < .001
Tokens per type	1.64 (0.27)	1.60 (0.18)	0.92(35)	-0.06 to 0.15	p = .365
Tokens per family	1.87 (0.32)	1.72 (0.23)	2.46(35)	0.03 to 0.27	p = .019
Characters per word	4.08 (0.36)	4.53 (0.24)	-6.11(35)	-0.59 to -0.30	p < .001
Syllables per word	1.23 (0.10)	1.36 (0.09)	-5.96(35)	-0.17 to -0.08	p < .001
Words per sentence	11.55 (3.54)	11.89 (5.89)	-0.36(35)	-2.28 to 1.59	p = .720
Flesch-Kincaid GL[c]	4.45 (1.33)	5.99 (1.50)	-5.67(35)	-2.09 to -0.99	p < .001
Coleman-Liau Index	7.87 (1.56)	10.93 (1.38)	-8.70(35)	-3.77 to -2.34	p < .001
New Dale-Chall	2.51 (0.79)	3.19 (0.80)	-3.53(35)	-1.08 to -0.29	p = .001
Vocab profile — k1[d]	87.3% (3.6%)	85.2% (6.9%)	1.81(35)	-0.26 to 4.57	p = .079
Vocab profile — k2[e]	6.9% (2.9%)	7.3% (4.1%)	-0.41(35)	-2.01 to 1.32	p = .680
Vocab profile — AWL[f]	0.8% (0.9%)	1.2% (1.5%)	-1.42(35)	-0.98 to 0.17	p = .164
Vocab profile — Offlist[g]	4.9% (2.4%)	6.3% (4.2%)	-1.81(35)	-2.98 to 0.17	p = .079

[a] SR = self-reference

[b] OR = other-reference

[c] Flesch-Kincaid Grade Level is a composite measure based on words per sentence and syllables per word.

[d] The k1 list consists of the 1,000 most common words in English.

[e] The k2 list consists of the second 1,000 most common words in English.

[f] The Academic Word List consists of 570 words that commonly occur in academic texts.

[g] Off-list words do not belong to the above three categories and thus tend to be less frequent.

more characters per word and more syllables per word. This suggests the use of longer lexical items, which are often more specific words with low frequencies of occurrence.

The vocabulary profile provided a calculation of the percentage of the text that was made up of (roughly) the first thousand word families, (roughly) the second thousand word families, the 570 word families of the Academic Word List (Coxhead, 2000), and "off-list" words (words beyond these three lists). Greater use of off-list words and the use of longer words are associated with the use of lower frequency words, the use of proper names and greater lexical sophistication.

The vocabulary profile measures yielded nonsignificant results. As mentioned in the Method section, the battery of measures reported here were not independent, so only the MLTU measure can be used to directly address the issue of complexity. The other measures have been reported for the sake of comparison.

Accuracy

Errors were calculated using an adapted version of the Foster and Wigglesworth (2016) formula ranging from 0 to 3, with 0 representing an error-free unit and 3 representing serious errors. Essays in the SR condition had lower error ratings ($M = 0.97$, $SD = 0.97$, range = 0.30-2.10) than those in the OR condition ($M = 1.19$, $SD = 0.97$, range = 0.54-2.21). At an alpha of .017, a paired-samples t-test showed that essays in the OR condition received significantly higher error ratings than those in the SR condition, $t(35) = 2.87$, $p = .007$, $d = 0.48$. The answer to the third research question would thereby appear to be negative: impersonal writing prompts are not associated with greater accuracy. In fact, the opposite is true: the SR prompts are associated with greater accuracy.

Discussion

The results suggest that the use of personalized (i.e., SR) writing prompts leads to more output and greater accuracy. Turning first to the findings for greater output, while the precise factors associated with increased output cannot be identified based on the current experimental design, it may be hypothesized that personalized prompts reduce cognitive load during the initial stage of generating ideas (part of the initial "planning" phase in the four-part model created by Hayes & Flower, 1980). Put simply, writers must simultaneously think of what to

say and how to say it. For the native speaker and highly proficient L2 writer, both of whom enjoy highly developed and automated linguistic resources, attentional resources can be easily shifted to the cognitive operations related to planning what to say. L2 learners, with less developed linguistic resources, do not have this luxury. Personalization, by drawing on linguistic resources that are (presumably) more developed and automated, may aid the linguistic encoding of these ideas (the "translating" phase in Hayes & Flower's 1980 model).

The current study failed to show that impersonal prompts are associated with more complex language, yet because the results fall just short of statistical significance (at a stringent alpha of .017), further research is warranted to determine how the experimental manipulation of personalization affects complexity in L2 writers' output. It is possible that more advanced learners, possessing a greater store of metalinguistic knowledge to draw on, would produce significantly more complex language in the OR condition. As Robinson (2015) suggests, cognitive complexity along resource-directing dimensions is often accompanied by complexity in the language employed to convey the complex concepts (for examples, see Robinson, 2015, p. 97). Impersonal prompts, by calling for a third-person narrative, may lead to more complexity in a number of subtle ways. To give just one speculative example, it could be that statements about other people tend to involve greater use of hedging and that this leads to longer and more complex syntax related to greater use of modals (e.g., *may*, *might*, *could*) and hedging phrases (e.g., there is a strong possibility that...). Moreover, OR prompts often require a third-person vantage point and overt third-person marking on verbs, which is inherently more difficult than first person (more commonly used in response to SR prompts) in which verbs appear in their bare forms. To develop a fuller account of the effects of personalization, future research will need to go beyond general measures such as those used here so as to examine the specific linguistic resources associated with the task and difficulty that learners at various levels experience when attempting to respond to OR prompts.

An important finding in the current research is that impersonal prompts are associated with an increase in error. If future research finds an association of OR

and greater complexity, this may show support for Skehan's (1998) contention that task complexity is broadly associated with more error. The findings would appear to run counter to the Cognition Hypothesis (Robinson, 2015). It should be noted, however, that Robinson acknowledges certain limiting conditions regarding his prediction for greater accuracy on monologic tasks that are more complex along resource-directing dimensions (in the current study, tasks involving other-reference). Specifically, he states that "increased accuracy and complexity on complex tasks will likely be found *most clearly* for those high in the abilities the complex task draws on, and may not be found for those low in these abilities" (p. 107, italics in the original).

This study has several pedagogical implications. Robinson (2015) has argued persuasively that pedagogical tasks need to be aligned with learners' current abilities and affective dispositions. Greater verbosity in replies to SR prompts would indirectly suggest that L2 writers find it easier to craft replies to these prompts. If further research confirms that this is the case, such prompts may be particularly useful when teaching students who experience especially high levels of anxiety when writing in an L2. SR prompts may also be more useful to offset the inherent difficulty of the task when students are asked to write about an unfamiliar topic.

The results do not imply that impersonal prompts should be avoided altogether. Pedagogical tasks, after all, must prepare students for real-world tasks, and part of real-world writing, especially in academic contexts, involves describing the actions and thoughts of other people or the assumption of an impersonal vantage point. OR tasks may also be justified on the grounds of L2 processing. If future research confirms that impersonal prompts lead to more complexity in output, such prompts may be useful ways to push students toward production that leads them to notice the gap (Swain, 1995) between their interlanguage resources and the message they intend to convey. Robinson's (2015) model of task sequencing would likewise suggest that altering the task toward a more complex dimension (in this case, going from SR to OR) would be necessary to promote a restructuring of the learner's current interlanguage system. For example, learners at the proficiency level of the participants often

fail to use plurals when making general statements about others (e.g., they say "a Japanese student studies hard" instead of "Japanese students study hard"). The OR task may make them more sensitive when listening or reading to patterns used by NSs when making such statements and thereby make it more likely that their interlanguage is restructured to include the target grammatical structure. Since impersonal prompts are associated with more errors, instructors may need to accompany writing tasks employing such prompts with greater scaffolding and should introduce the tasks later in the task cycle.

The current research has reached some tentative conclusions, but more work is needed to determine the effects of personalization for different populations, contexts, and modes (e.g., writing versus speaking). It could be the case that personalized prompts are especially important for certain groups (e.g., lower-proficiency learners or learners with more anxiety). The effects of personalization on spoken tasks, which generally entail more time pressure, may also be different from those reported in the current study. In addition, future research must enlist appropriate theoretical frameworks, ranging from general psychological research to SLA and L2 writing research, to develop a more comprehensive understanding of the effects of personalization on L2 learning. It should be noted that research based on task complexity (e.g., Robinson, 2015) during the last two decades has focused predominantly on speaking. This has led to a large body of research with comparable results, but it has meant that the application of task-based theories of instructional design to writing has been relatively neglected.

Finally, continued research on personalized tasks will hopefully provide avenues for cross-pollination between several broad research agendas. On the one hand, SLA theory construction could benefit from greater integration of memory research that has examined the self-reference effect (e.g., Bellezza, 1984; Symons & Johnson, 1997) and autobiographical memory (e.g., Conway, 2005; Conway & Holmes, 2004). Looking forward, writing pedagogy is also likely to be increasingly influenced by the broad research agenda examining the self as it relates to emotion, agency, autonomy, self-esteem, motivation, and other related constructs (e.g., MacIntyre, Gregersen, & Mercer, 2016; Mercer & Williams, 2014).

References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam, The Netherlands: John Benjamins.

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.

Chall, J. S., & Dale, E. (1995). *Manual for use of the new Dale-Chall readability formula*. Brookline, MA: Brookline Books.

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.

Cobb, T. *Web Vocabprofile*. [Computer program]. Retrieved from http://www.lextutor.ca/vp/eng/

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283-284.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594-628.

Conway, M. A., & Holmes, A. (2004). Psychosocial stages and the availability of autobiographical memories. *Journal of Personality*, *72*(3), 461-480.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387.

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.

Fujii, Y. (2012). The effects of L1 and L2 writing and translation: A case study. *Journal of Modern Languages*, *22*, 32-44.

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

Harris, C. B., Rasmussen, A. S., & Berntsen, D. (2013). The functions of autobiographical memory: An integrative approach. *Memory*, *22*(5), 559-581.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Heatley, A., & Nation, P. (2002). *Range* [Computer program]. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation#vocab-programs

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). Amsterdam, The Netherlands: John Benjamins.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. *NCTE Research Report No. 3*. Champaign, IL: National Council of Teachers of English.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the society for research in child development*, 35(1), iii-67.

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, *10*(3), 2-10.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel (Research Branch Report 8-75)*. Millington, TN: Naval Technical Training, U.S. Naval Air Station, Memphis, TN.

Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 117-135). Clevedon, UK: Multilingual Matters.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48-60.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking: The effect of mode. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 91-104). Amsterdam, The Netherlands: John Benjamins.

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge.

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

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

MacIntyre, A., Gregersen, T., & Mercer, S. (Eds.). (2016). *Positive psychology in SLA*. Bristol, England: 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.

Mercer, S., & Williams, M. (Eds.). (2014). *Multiple perspectives on the self in SLA*. Bristol, England: Multilingual Matters.

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.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*(4), 218-233.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492-518.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, *47*(1), 101-143.

Révész, A. J., Kourtali, N.-E., & Mazgutova, D. (2016). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language*

Learning, *67*(1), 208-241. doi: 10.1111/lang.12205

Richards, J. C. (2006). Materials development and research—Making the connection. *RELC Journal*, *37*(1), 5-26.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge, England: Cambridge University Press.

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.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. P. Garcia-Mayo (Ed.), *Investigating tasks in formal language settings* (pp. 7-26). Clevedon, England: Multilingual Matters.

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). Amsterdam, The Netherlands: John Benjamins.

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.

Sasaki, M. (2011). Effects of varying lengths of study-abroad experiences on Japanese EFL students' L2 writing ability and motivation: A longitudinal study. *TESOL Quarterly*, *45*(1), 81-105.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, UK: Cambridge University Press.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165-202.

Segalowitz, N. S. (2003). Automaticity and second languages. In C. J. Doughty

& M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 382-408). Malden, MA: Blackwell.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Slobin, D. I. (1993). Adult language acquisition: A view from child language study. In C. Perdue (Ed.), *Adult language acquisition: Crosslinguistic perspectives, Vol. 2: The results* (pp. 239-252). Cambridge, England: Cambridge University.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honor of H. G. Widdowson* (pp. 125-144). Oxford, England: Oxford University.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371-394.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning*, *52*(4), 755-803.

Way, D. P., Joiner, E., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, *84*(2), 171-184.

Author bios

Charles Mueller *received his Ph.D. in Second Language Acquisition (SLA) at the University of Maryland. He currently teaches courses on SLA, cognitive linguistics, and academic English at Fuji Women's University in Sapporo, Japan. His research has been primarily conducted within the usage-based theoretical approach. mueller@ fujijoshi.ac.jp*

William Kraus *received an MA in Ancient History from Columbia University and an MA TESOL from the Middlebury Institute of International Studies. He has taught ESL and EFL at American and Japanese universities. Since 2007, he has*

worked at Fuji Women's University as coordinator for first- and second-year English programs. billkraus9@yahoo.com