VU Research Portal

Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts

de Jong, C.A.M.; Vercellotti, M.L.

published in

Language Teaching Research
2016

DOI (link to publisher)

[10.1177/1362168815606161](https://doi.org/10.1177/1362168815606161)

document version

Peer reviewed version

[Link to publication in VU Research Portal](#)

citation for published version (APA)

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

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

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

<https://doi.org/10.1177/1362168815606161>

Similar Prompts May Not Be Similar in the Performance They Elicit: Examining Fluency, Complexity, Accuracy, and Lexis in Narratives from Five Picture Prompts

Interest in task-based learning, with entire books (e.g., Ellis, 2003; García Mayo, 2007) devoted to the issue, has increased focus on the effect of task on language performance. Studies like those by Skehan and Foster (1997) have specifically compared narrative tasks, information sharing tasks, and decision-making tasks to see how task influences language performance. Often, however, researchers have focused on a single type of language task, such as picture narration or topic response, studying task complexity by manipulating tasks within that genre (see Révész, 2014). For instance, Michel, Kuiken and Vedder (2007) asked learners to describe a decision based on only two choices (simple condition) or six choices (complex condition), while Robinson (1995) had learners speak about a comic either in the present while looking at the comic (simple condition) or in the past without looking at the comic (complex condition.) This exploration of task complexity is relevant as it gives insight into the cognitive demands of the task, thereby informing language learning theories, pedagogy, and assessment.

Although researchers, language teachers, and test developers have often elicited narratives with picture story (or comic strip) prompts, relatively little is known about what

FIVE PICTURE STORY PROMPTS

features of the story prompts influence task performance, to what extent, and why. As a result, it is not easy to identify equivalent prompts for parallel tests or to select prompts that differ only in the feature under investigation. In this paper, we show that picture prompts that seemed *a priori* to be equivalent in terms of three known complexity dimensions nevertheless elicited variable performances in fluency and lexis. We explore how the prompts might differ in additional dimensions of the task design and discuss a number of important consequences for research on task demands.

Task complexity and linguistic performance

Second language (L2) learning theory and research have considered the effect of task demands on language performance. Skehan (2009) and Robinson (2011) linked task demands to attentional resources, and the subsequent effects on fluency, accuracy, and complexity of the performance. The influences on task performance include both the conditions under which the tasks are performed and the characteristics of the tasks themselves. Several task demands have been identified by Skehan and Robinson, including quantity of information and reasoning demands. In Robinson's model, task complexity includes cognitive factors that are "resource-directing" (e.g., $\pm$ elements, $\pm$ causal reasoning) and "resource-dispersing variables" (e.g., $\pm$ task structure, $\pm$ prior knowledge). In addition to cognitive factors, Robinson's model includesⁱ learner factors (e.g., motivation, anxiety) which affect performance on language tasks. According to Robinson's Cognition Hypothesis, more complex tasks along the resource-directing dimension can elicit language that is more complex and more accurate, and more complex tasks in resource-dispersing dimensions can lower performance.

FIVE PICTURE STORY PROMPTS

In Skehan's Extended Trade-off Hypothesis (Skehan & Foster, 2012) task type and task complexity are likely to impact language performance. For instance, narratives tend to have higher lexical variety, often at the expense of fluency and accuracy, according to Skehan. But, Skehan (2009), unlike Robinson, argued that when second language speakers are pressured to produce more complex language, attentional capacity is used for conceptualization and will therefore be reduced for subsequent stages like lexical retrieval and syntactic structure building, resulting in lower accuracy and/or fluency. Nevertheless, it may be possible for speakers to direct their attention to both complexity and accuracy in the same task. For Skehan, this is possible because these performance aspects are raised by different factors.

Task complexity can be manipulated with features of the task by changing the instructions or the conditions given to the participants. For instance, Elder, Iwashita, and McNamara (2002), manipulated task complexity by making a narrative speaking task less or more "difficult" in four dimensions (perspective, immediacy, adequacy, planning). And in Kuiken and Vedder's (2008) writing task about choosing a vacation destination, the participants in the complex condition had to make the decision considering more requirements than the simple condition. The results showed that the more complex tasks (+elements) elicited increased accuracy but had no influence on grammatical complexity or (strangely) lexical variety (see also Michel, Kuiken & Vedder, 2007). But when participants performed similar decision-making tasks in different planning and $\pm$ elements conditions, Levkina & Gilabert (2012) found only some impact on fluency and lexis, depending on planning condition, and no impact of $\pm$ elements on accuracy or complexity.

FIVE PICTURE STORY PROMPTS

In such studies, research design and methodology are crucial to interpreting the results. Different picture prompts are needed to elicit the narratives for the different conditions (e.g., Rahimpour, 2010; Ishikawa, 2007) to avoid repetition effects. The use of single prompts for each condition, however, creates a confound between task condition and task prompt. For instance, based on a comparison of only two prompts Skehan and Foster (1999) concluded that predictability of the storyline affected fluency, in that more predictable stories elicit more fluent language performance, but that conclusion could have been skewed by unknown prompt effects. One option to mitigate any unknown prompt effects is to counterbalance the prompts with task condition (e.g., Robinson, 1995) or to use parallel forms. Although this is possible for decision-making tasks, it is much more difficult for picture narration tasks. One study (Bae & Lee, 2010) confirmed that two picture prompts were parallel based on raters' scoring (0-4) of the elicited texts. Another possible, but rare (or unreported), solution to address the confound between condition and prompt is piloting of the prompts by empirically comparing the elicited language (cf. Ahmadian, Tavakoli, and Dastjerdi, 2013). More commonly, however, researchers have attempted to control for prompt differences in prompt selection. For instance, Elder, Iwashita, and McNamara (2002: p. 354) reported that "the two exemplars in each dimension resembled each other as closely as possible in terms of their linguistic demands and likely familiarly to the test-takers." Likewise, Kormos and Trebits (2012: p. 450) stated that the prompts in their cartoon narration task were "selected carefully to include similar elements in the two versions" and that two prompts in cartoon description task had "the same number of actors and key narrative events". Similarly, albeit more generally, van Hest (1996: 31) stated that the sets of prompts "had to be similar in length and degree of difficulty". Researchers' pre-

FIVE PICTURE STORY PROMPTS

use evaluation of prompt similarity, however, may be unreliable and not reported in sufficient detail, and the impact on the elicited language performance is unknown.

Three Task Characteristics of Narratives

Only a few studies have considered how language performance is affected by specific features in narratives (e.g., Ahmadian, Abdolrezapour, & Ketabi, 2012; Tavakoli & Foster, 2008). One narrative feature relates to the connectedness of the picture frames. Picture stories with a ‘tight’ inherent narrative structure have a clear time sequence from beginning, middle, and end, while in a ‘loose’ structure “the events could be reordered without compromising the story” (Tavakoli & Foster, 2008, p. 447). In Robinson’s (2011) Cognition Hypothesis, this dimension might be considered $\pm$ causal reasoning, since narratives with a “loose” structure require participants to create the connections between frames whereas narratives with a “tight” inherent structure have more obvious connections. Ahmadian, Abdolrezapour, and Ketabi (2012) reported that stories with a “tight” narrative structure allowed speakers to focus on grammatical or lexical repairs, which was attributed to more attentional resources for monitoring, and that learners focused on more meaning-based repairs during the task with a “loose” narrative structure. Comparing the language of narratives elicited from pictures already in a coherent structure and narratives created from unrelated pictures, Kormos & Trebits (2012) found increased lexical diversity in the tight structure, which was attributed to less flexibility, and no difference in complexity, attributed to the resource-dispersing nature of oral performance.

Tavakoli and Foster (2008) and Tavakoli (2009) examined how narrative structure and storyline complexity may impact performance. Storyline complexity refers to the presence of

FIVE PICTURE STORY PROMPTS

background information in addition to foreground information. Foregrounded events are those that tell the story, whereas background information elaborates on, explains or evaluates the foreground events. Tavakoli and Foster studied the language elicited by four picture prompts, finding that the story prompt identified as having a complex storyline and a tight narrative structure had the highest means for complexity, accuracy and fluency. They concluded that a complex narrative storyline increased syntactic complexity (measured by a subordination ratio, and clauses/AS-unit), while a tight narrative structure (clear, fixed sequence) helped non-native fluency. Likewise, Tavakoli (2009) studied the effects of narrative structure and storyline complexity with six picture prompts and found that the presence of background information affected syntactic complexity, raising syntactic complexity for all narrative structures. For fluency and accuracy, on the other hand, Tavakoli reported mixed effects of task structure. These findings were seen as support for Skehan's (2009) claim that task complexity mostly affects conceptualization and therefore complexity of the language produced, while narrative structure affects processes in the formulator stage and therefore mostly affects fluency (and accuracy).

Caution is necessary, however, when interpreting the results from these two studies (i.e., Tavakoli & Foster, 2008 and Tavakoli, 2009), because only one prompt represented each combination of storyline complexity and narrative structure. The prompts chosen to represent the dimensions might not have been equivalent in other aspects than the two features under consideration. In fact, in a follow-up study, Tavakoli (2011) described differences between native and L2 learners in the number of end-clause pauses in only two of four picture-prompts, without an exploration of why the prompts may have elicited different language performances.

FIVE PICTURE STORY PROMPTS

Similarly, Tavakoli and Foster (p. 457) reported an “inconsistent effect” in lexical variety (D) between the two prompts labeled “loosely structured”. A plausible explanation is that another dimension of complexity is at play, and a conspicuous one is the number of characters or props, which could be considered $\pm$ elements in Robinson’s Cognition Hypothesis. It is likely that a prompt with a variety of characters or props will have a higher D score because the prompt requires the speaker to introduce new lexical items, whereas for prompts with fewer characters and props, participants (rightly) recycle those lexical items, resulting in a lower D score. In fact, $\pm$ elements has been considered in task complexity research (e.g., Kuiken & Vedder, 2008) and controlled for in others (e.g., Kormos & Trebits, 2012).

In sum, research has shown performances differ when elicited by the different prompts despite their pre-determined assumed equivalence. The present study can be considered a follow-up to the studies reported by Tavakoli and Foster (2008) and Tavakoli (2009; 2011), although it was not originally conceived as such. It investigates whether prompts *within* certain categories elicit similar performances. In this exploratory study, all of the prompts had a tight sequential structure (-causal reasoning), similar storyline complexity (+intentional reasoning), and similar number of main characters and props (-elements), three often cited and salient narrative features.

Purpose of the study

The main purpose of this study was to examine whether prompts with similar narrative structure, storyline complexity, and number of elements elicit similar performance in terms of fluency, complexity, lexis, and accuracy. If there are differences, despite controlling for these features, we explored what might account for these differences. Possible influences included

FIVE PICTURE STORY PROMPTS

other features of the story, such as the similarity of the characters, trade-offs between complexity, accuracy, fluency, and lexis during language performance, and differences in task difficulty.

II Method

a Participants

Participants were adult ($M = 25.8$ years) English L2 learners in high-intermediateⁱⁱ speaking classes in an intensive English program at an American university ($n = 25$). They had mixed language backgrounds: Arabic (4), Chinese (8), Japanese (3), Korean (5), and single speakers of Samoan, Spanish, Taiwanese, Thai, and Turkish. The mixed language backgrounds added variability but also mitigated specific L1 effects.

b Materials

Informed by Rossiter et al.'s (2008) guidelines, we reviewed ESL teaching materials, SLA research, and L1 research for potential six-frame picture prompts for a “cartoon description” (Kormos & Trebits, 2012) speaking task. This set was then narrowed down after feedback from ESL teachers and native speakers. From Heaton (1966), three six-frame picture-prompts were selected (Bicycle, Race, and Tiger). Two additional six-frame prompts (Frog, Turtle) were created by choosing pages of wordless picture books (Mayer, 1967; Mayer & Mayer, 1971), so as to test a wider variety of prompts. Heaton and Mayer prompts have been used in other bilingual research (e.g., Kormos & Trebits, 2012; van Hest, 1996).

FIVE PICTURE STORY PROMPTS

All stories had a tight sequential structure, (i.e., the frames cannot be reordered without compromising the story), which is -causal reasoning in Robinson's (2011) Cognition Hypothesis and a complex storyline with background information. Specifically, each prompt encouraged participants to discuss the motivation or goals of the characters in the story (+intentional reasoning), and each prompt had an element of surprise. In addition, each of the prompts had two main characters and only two locations (-elements) so that they would require about the same level of description, a feature considered in other studies (e.g., Kormos & Trebits, 2012).

Because we were interested in studying the participants' best language performance (as part of a larger study), guiding questions accompanied each picture prompt to provide the participants with additional suggestions for content, potentially reducing individual differences. Some of the questions were suggested within Heaton (1966). All prompts and guiding questions are available in the IRIS system.

c Procedures

During the regular speaking class, prompts were displayed and narratives were recorded on Dell computers using a software program developed with Runtime Revolution (Shafer, 2006). For each prompt, participants were given three minutes planning time and four minutes to tell the story. During the planning time, the participants saw the pictures and guiding questions and could zoom in to individual pictures by clicking on them. During the recording, only the static picture prompts were visible.

The narrations were collected in two class sessions, three per session. Participants were randomly assigned to one of two orders of presentation to minimize order effects. All

FIVE PICTURE STORY PROMPTS

participants saw the same practice prompt (Picnic, also from Heaton, 1966) for the first narration of the first session, and the order and scheduling allowed four different prompts to be presented last (and thus subject to fatigue) on data collection days. Order A presented Picnic, Tiger and Turtle on the first day and Race, Bicycle, and Frog on the second; Order B presented Picnic, Frog, Bicycle and then Turtle Tiger, Race. After each narration, participants provided information about their experience, such as reporting any lexical difficulties.

d Measures and Analyses

The prompts were evaluated by several measures of language performance: fluency, lexical variety, complexity, and accuracy (see Table 1). Only data from those participants who completed both sessions (gave narratives for all five prompts) were analyzed. Each participant thus serves as her own control in the within-subject design, mitigating effects of individual differences.

[TABLE 1 HERE]

The recordings were trimmed to the point where the participant stopped talking or took a long pause, indicating that the participant had finished the narration. A “long” pause was considerably longer than earlier pauses, in all cases longer than nine seconds. Admittedly, any cut-off point is arbitrary, but the cut-off was made at a clear gap in the participants’ pause lengths. The mean amount of speech produced was 218.3 seconds ($SD=29.3$) for Bicycle, 222.9 seconds ($SD=21.8$) for Tiger, 212.7 seconds ($SD=40.9$) for Race, 217.8 seconds ($SD=29.7$) for

FIVE PICTURE STORY PROMPTS

Frog, and 205.1 seconds ($SD=53.4$) for Turtle. There were no significant differences between the prompts for amount of speech produced.

Fluency was evaluated by three temporal measures (mean length of fluent run, mean pause length, and articulation rate) following the computations used by Author1 and Coauthor1 (2011) and two measures of repair (number of repeated words, number of reformulated words). Pauses were defined as silent or filled (e.g., *uh*, *mmm*) pauses of 200ms or longer.

Three indicators related to lexis were examined. First, lexical variety was calculated using a lemma-based version of D , which is a type-token ratio measure that controls for text length (McKee, Malvern, & Richards, 2000). Because D is a text-internal measure, taking into account only those words actually used in text, it is less affected by topic effects than a text-external measure. To explore topic effects in more detail, our second indicator of lexical performance was a text-external measure, which takes into account potential words that were not used. Text-external measures (e.g., P_Lex) are generally based on a predetermined frequency list, but it is difficult to choose an appropriate word list for a particular population or task. Importantly, Meara and Bell (2001) suggested that an improvement on the P_Lex measure would be to create “task-specific word lists”. Although they suggested using native-speaker data, it might be more applicable to use task-specific and population-specific words. As Granger (2003, p. 543) stressed “native corpus data, however detailed, will never tell anything about the difficulty of words and structures for learners.” Therefore, we generated lists of words that were most commonly used by the participants for each prompt, creating lists which were specific to the population as well as the task. The common words were identified as those which were used by twelve or more participants (about half of the sample population) within a

FIVE PICTURE STORY PROMPTS

prompt. These lists were then examined for similarities and differences across prompts. The third indicator of lexical issues concerned potential lexical retrieval difficulties. After each narrative recording, participants were asked if there were any words they would have liked to use but did not know or could not remember; they could answer in English or in their native language. The prompts were then compared with respect to the number and type of words reported.

Next, complexity was evaluated with three measures, following Norris & Ortega (2009), based on AS-units (Foster, Tonkyn, & Wigglesworth, 2000): global complexity (number of words per AS-unit), phrasal complexity (number of words per clause), and complexity by subordination (number of clauses per AS-unit). These three measures capture compositional complexity as opposed to “structural complexity”, which considers relationships between constituents (Rescher, 1998; Vercellotti, 2015).

Finally, accuracy was measured at two levels of production (error-free clauses and error-free AS-units) and by the number of corrections. Although corrections can disrupt fluency, they show an orientation toward accuracy (Ellis & Barkhuizen, 2005).

All performance measures were analyzed using general linear model ANOVA with story prompt as the within-subject factor. Since there was no effect of order of presentation or interaction between order and prompt, the analyses did not use order of presentation as a factor. Whenever the assumption of sphericity was violated, the Greenhouse-Geisser adjustment was used. Whenever significant differences were found, post-hoc pairwise comparisons using the Bonferroni adjustment were made.

III Results

a Fluency

Fluency was evaluated by five different measures (see Table 2). There was a main effect of prompt on mean pause length, $F(2.357, 56.557) = 3.450$, $p = 0.032$, $\eta^2_p = .126$, with the Turtle narratives having longer pauses than the Tiger narratives ($p = .018$). In addition, there was a marginally significant effect of prompt on the number of repetitions, $F(2.445, 58.682) = 2.744$, $p = .062$, $\eta^2_p = .103$. Frog had fewer repetitions per 100 words than Bicycle ($p = .030$) and Race ($p = .045$); no other comparisons reached significance. Mean length of fluent run, articulation rate, and reformulated words did not show a significant effect of prompt.

[TABLE 2 HERE]

b Lexis

Lexical variety, measured by D on lemmas (Table 3), provided a text-internal measure of lexical performance. One participant's scores were deleted because her D score on the Turtle narration was an outlier (more than three *SDs* from the mean). There was a main effect for prompt, $F(4, 92) = 17.884$, $p < .001$, $\eta^2_p = .437$. Post-hoc analyses showed that Race elicited narratives with significantly higher lexical variety compared to Bicycle ($p = .002$), Frog ($p < .001$) and Turtle ($p < .001$). Also, Frog elicited significantly less lexical variety than Bicycle, ($p = .048$), Tiger ($p < .001$), and Race, ($p < .001$), and a trend with Turtle ($p = .091$). No other comparisons revealed significant differences.

FIVE PICTURE STORY PROMPTS

[TABLE 3 HERE]

As a text-external measure of lexical performance, we created a task- and population-specific frequency word list, calculated by the number of participants who used it. We compared the lists across prompts, reviewing use of function words and content words (nouns, verbs, adjectives, and adverbs). All five prompts—as expected—most commonly elicited the function words *and*, *the*, and copula *be*. Infinitival *to*, *a*, *he*, and auxiliary *be* were also used by most participants for every prompt. The prepositions *in*, *of*, and *to* were common in each prompt’s narratives (16 – 24 participants per prompt). For each prompt, the two most common subordinate conjunctions were *when* and *because* (14 – 19 participants). But there were also some differences between prompts. One result of interest is that Race was the only prompt to elicit three auxiliaries in 12 or more participants: auxiliary *be* (20), *will* (14), and *can* (12). Also, Tiger seems to have elicited the existential use of *there* (as in the sentence structure “there is/are”) most often, with 23 participants using that construction at least once, followed by Race (18), Bicycle (11), Frog (11), and Turtle (6).

Compared to function words, content words are more likely to differ among prompts. Table 4 lists the content lemmas that were used by 12 or more participants for each narrative. The main characters, objects, and actions are well represented in the top ten, but the results may be more interesting for what is *not* there, rather than what is there. For each prompt, some seemingly useful lemmas were not used by the majority of participants. In general, two observations stand out. First, Race had fewer shared lemmas, specifically fewer nouns used by all or most participants, unlike the other prompts where nouns dominate the top five in

FIVE PICTURE STORY PROMPTS

frequency. It is striking that no main characters or objects occur in the top 10, only a number of key actions (*run, race, win*) and one attribute of a main character (*tall*). In contrast, the main characters all occur in the top 5 for the other prompts. A closer look at the Race narratives found that the main characters were referred to in a variety of ways: *boy, guy, man, person, number* (e.g., *number one*), *participant, child*, and *character*. A second striking observation was that in the Turtle narratives, *turtle* was not used by five participants, even though that word seems essential to its story (also see Discussion). These five participants used the more generic nouns *animal* or *fish* to refer to the title character. Taken together, the results of this analysis show that there is considerable variability in the content lexis elicited by the different prompts, with only a limited number of words elicited in a majority of participants.

[TABLE 4 HERE]

Finally, as an indication of lexical retrieval difficulties, we examined the words reported by the participants as unknown or not remembered during the task. Fifteen participants reported that they did not know or could not retrieve certain words that they wanted to use. These unknown words included 38 words or phrases, eight for each prompt, except Tiger with six, so the number of reported difficulties were similar among the prompts. From the given descriptions, the following words or concepts were interpreted to be reported as unknown:

Bicycle: *horn, impatience, ringing a bell, pass, fall, fail or breakdown, "serves you right"*

Tiger: *unconscious, hunter, walking stick, tiger, trail, kill with a stone*

Race: *race/marathon, run a race, brag, number bib* (2 participants)

FIVE PICTURE STORY PROMPTS

Frog: *net* (3), *escape* (2), *corner/surround the frog* (2), *bathroom/bathtub*

Turtle: *turtle* (3), *bury/funeral*, *cemetery/grave*, *fishing rod* (2), *floating*

For three comments concerning Race, it was not clear what the participant was reporting as unknown (*YR*, *exceed*, *claver* or *meening* (sic) *about something*). One additional comment from Bicycle seems to refer to a situation without a corresponding English word or phrase, *feeling to win for someone who you don't like*.

Overall, the combination of the three analyses (text-internal, text-external, and familiarity) provides a detailed overview of the lexical performance elicited by the prompts. The relatively high lexical variety (D) in the narratives elicited by Race could be explained by the varied ways in which the characters were referred to, as evidenced by the lack of commonly used words for these characters found in the text-external measure. Frog elicited a relatively low lexical variety but did not elicit a greater number of commonly used words than the other stories. Another finding, from the word lists and familiarity reports, was that words for main characters were not known and not used by all speakers for all stories: *tiger* was reported unknown by one participant and not used by two, and, *turtle* was reported unknown by three participants and not used by five. (Because some referred to the characters in Bicycle by a proper noun, *boy* or *man* were not always used by all participants.)

c Compositional complexity

Complexity of the narratives was measured in global complexity (words per AS-unit), phrasal complexity (words per clause), and complexity by subordination (clauses per AS-unit) (Table 5). The effect of prompt was not significant for any of the measures, $F(4, 96) = 0.63$, $p = .834$, $\eta^2_p =$

FIVE PICTURE STORY PROMPTS

.015; $F(4, 96) = 2.039$, $p = .095$, $\eta^2_p = .078$, and $F(3.024, 72.588) = .324$, $p = .810$, $\eta^2_p = .013$, respectively. The slight trend for phrasal complexity is driven by the difference between Bicycle and Race ($p = .131$). All prompts seemed, therefore, to elicit performances of comparable compositional complexity.

[TABLE 5 HERE]

d Accuracy

Accuracy was measured as percentage of error-free clauses, percentage of error-free AS-units, and number of corrections/100 words (Table 6). There was no effect of prompt for accuracy scores as measured by percentage of error-free clauses, $F(4, 96) = 2.182$, $p = .077$, $\eta^2_p = .083$, although there was a trend; the post-hoc analysis showed that Frog had a somewhat higher error-free clause percentage than Bicycle, $p = .088$. There was no effect of prompt for the accuracy scores based on error-free AS-units, $F(4, 96) = 1.073$, $p = .374$, $\eta^2_p = .043$. Finally, although the mean number of corrections per 100 words differed greatly from prompt to prompt, the differences were not statistically significant, $F(2.627, 63.037) = 1.630$, $p = .196$, $\eta^2_p = .064$. There was no observed difference between the prompts in terms of the accuracy of performance elicited.

[TABLE 6 HERE]

FIVE PICTURE STORY PROMPTS

In summary, we observed a number of differences in the narratives elicited from these prompts, which were *a priori* expected to be similar. Turtle elicited less fluent performance (as measured by mean pause length) than Tiger. Frog induced the lowest lexical variety and Race the highest. Notably, five out of 25 participants did not use the lemma *turtle* for the Turtle story (see Discussion). There were no significant differences in any of the measures of compositional complexity and accuracy.

IV Discussion

The goal of this study was to determine whether L2 performance was similar across prompts that were alike in narrative structure, storyline complexity and number of elements.

Performance was indeed similar in terms of complexity and accuracy. Since previous research testing the Cognition Hypothesis (with non-narrative tasks) found an increase in accuracy for more complex tasks (Kuiken & Vedder, 2008; Michel, Kuiken, & Vedder, 2007), this might indicate that the chosen prompts created equally complex tasks. There were, however, differences in fluency. Fluency was lower for Turtle than Tiger, while the fluency measures of Bicycle, Race, and Frog were not statistically different from the other two prompts. Notably, the high standard deviations on the reformulation measure point to high between-individual differences; some participants regularly abandoned utterances while others did not use that strategy at all.

Differences were also found with respect to lexical performance, which was examined with a task-internal measure (lexical variety), a task-external indicator (word lists), and participants' reports of unfamiliar words. Lexical variety, as measured by D, showed substantial

FIVE PICTURE STORY PROMPTS

differences between the prompts. This contrasts with findings by Kuiken and Vedder (2008). The texts analyzed in their study were aided by dictionaries, whereas the oral language produced in the present study reflects the participants' lexical performance without reference materials or the benefit of reflection during writing. The analyses showed that the Race prompt elicited greater lexical variety. This lexical variety finding is interesting because the prompts were similar in the $\pm$ elements dimension. An explanation of the higher D score in the Race narratives may be found in the word list results, in that individual participants tended to refer to the main characters in several different ways. Consequently, and in stark contrast to the other four narratives, no references to the main characters of Race were among its top 10 of commonly used words. These lexical differences, however, were not reflected in the temporal measures of fluency, in that greater lexical variety did not seem to be related to higher or lower fluency across prompts. Accordingly, there is little support for a trade-off effect between lexical variety and fluency, despite assumed lexical retrieval costs. Each prompt elicited about equal numbers of comments about unknown words. Interestingly, some of these words seem essential to the story. For example, three participants reported not knowing *turtle*, and in fact five participants did not use that lemma in their narratives.

Overall, these results indicate that there were differences among the picture-prompt-based narratives and that the differences were not uniform. Tiger and Bicycle had high or average scores on all measures (fluency, accuracy, complexity, and lexis). And while Frog had low lexical variety, it had average fluency measures and few repetitions. On the other hand, Race had high fluency and high lexical variety, but it also seemed to induce many repetitions and variable references to the main characters. Of the five prompts, Turtle seems to stand out

FIVE PICTURE STORY PROMPTS

with low fluency, vocabulary difficulties, and less speech, all indicating that this prompt created a more difficult task for these participants. In sum, despite their apparent equivalence in narrative structure, storyline complexity, and number of elements, the prompts did not elicit similar performance in all measures.

These differences have serious consequences for L2 research, especially for studies comparing certain task characteristics. If only one prompt is used per condition, it is not clear whether it is an adequate representative of the combination of task characteristics or whether there are other, unknown, characteristics that affect performance. As argued, classifying picture-based prompts is difficult, and differences in performance can be driven by dimensions unexpected by the researcher. For stronger conclusions that a particular dimension is driving differences in performance, it is preferable to use more than one prompt per cell. In addition, researchers should describe and compare the prompts in the study in as much detail as possible. It is important to realize that if prompts are used for assessment, care needs to be taken to ensure that the prompts are as similar as possible for different groups or at different test sessions. If possible, prompts need to be piloted to ensure their similarity in terms of the performance they elicit.

It should also be noted that, when comparing across prompts, researchers should confirm their classification with the language produced by the participants. In other words, even if the stories seem similar to researchers, the participants may not treat them as such. For instance, if a story prompt is expected to have a background story, but the majority of the participants do not address that secondary story, whose perception classifies the prompt? Any classifying or explanatory features (e.g., complex storyline) should be verified in the language

FIVE PICTURE STORY PROMPTS

performance of the participants and reported as support for the prompt's classification. We must be careful, however, to avoid circular reasoning when identifying more difficult tasks solely by performance measures (Levkina & Gilabert, 2012); participants' comments, like those about unfamiliar words, are valuable data. Moreover, as Robinson (2003) pointed out, the dimension of task difficulty, based on the learners' perception can affect performance, but it is unlikely to be realized until after the data collection.

a Some tentative explanations

Since this study was exploratory in nature, only tentative explanations can be offered as to the causes of the observed differences. For the fluency differences, one possible cause may have been the cost of lemma retrieval, particularly the need to use less-frequent vocabulary (Skehan, 2009). To explore this idea, we re-analyzed data of the Turtle prompt, for which five participants did not use the central word *turtle*. The data of those five participants were reanalyzed, focusing on the three measures for which significant differences between the prompts were found: number of words, average pause length, and D. For some participants the Turtle narrative was an entire standard deviation from their average score on each measure (see Table 7). For example, for four of the five participants, their mean pause length for Turtle was at least one standard deviation longer than their average pause length over all five narratives, indicating that these participants seemed to have responded to the lack of *turtle* by using longer pauses. It is also important to note that these participants were not below average generally; each of their mean scores across prompts was higher than the group average for each of the fluency measures. Only participant 991 was less successful than the group and

FIVE PICTURE STORY PROMPTS

performed even worse during the Turtle prompt. The amount of speech and the lexical variety (D) findings were robust when excluding the five participants that did not use *turtle*, but the fluency difference was no longer significant. Although the power decreased somewhat in this analysis with fewer participants, the effect sizes also decreased (from partial $\eta^2 = .126$ to $.107$).

[TABLE 7 HERE]

It seems therefore that lexical retrieval *difficulties* (rather than retrieval of varied lexical items), for both higher-performing participants and a lower-performing participant, negatively impacted performance. Narratives based on picture prompts allow relatively little flexibility in what the participant can and should say, which Skehan (2009) called the non-negotiability of the task. Although these participants could have quickly adopted a term to consistently refer to the turtle, they did not extensively use substitution or circumlocution for the unknown words for characters and events in the stories. If they had, fluency and compositional complexity measures may have been similar to—or even better than—the other narratives (e.g., *the fish with a shell* would result in a longer clause than *the turtle*). The results support Skehan's assertion that the need to use less-frequent lexis and the non-negotiability of the task can cause performance pressure that disrupts parallel processing, i.e., the speaker cannot smoothly deliver the current utterance and simultaneously plan the next utterance.

Alternatively (or additionally), the Turtle prompt may differ from the other prompts in another dimension of task complexity. Each prompt required the participants to understand and discuss the intentions of the characters, but the participants may have had more difficulty

FIVE PICTURE STORY PROMPTS

explaining this story, such as the boy's intention to bury the frog. In Robinson's (2011) Cognition Hypothesis, $\pm$ intentional reasoning is a *resource-directing* feature, but a speaker's confusion regarding a character's actions and goals could disperse attentional resources. It may be that the turtle narrative perhaps differs in task *difficulty*. Robinson has explained that task difficulty from learner factors contributes to between-learner variation. Learner factors are separated into ability variables and affective variables. One ability variable is vocabulary knowledge. The evidence suggests that a few participants found this prompt more difficult because they did not know the word *turtle*, and their performance for the Turtle narrative was noticeably lower than for the other prompts. Affective variables include anxiety, confidence and interest in the task. In fact, this story was reported to be generally less interesting than the other stories and significantly less enjoyable to tell than all other stories, which may have degraded their performance. Robinson (2003) made two germane points regarding task difficulty. First, although task difficulty will likely have less influence on L2 performance than task complexity, fluency is one area that is more related to task difficulty. And second, task difficulty is not easy to determine beforehand. The results for the Turtle prompt support both points about task difficulty.

The accuracy results in the present study seemed to be unaffected by any prompt differences. These results echo Levkina & Gilabert's (2012) decision-making task findings regarding task complexity and $\pm$ elements. Likewise, all prompts performed similarly in compositional complexity. This was expected, because the prompts were similar in narrative structure and storyline complexity. It may be that measures for these constructs, particularly

FIVE PICTURE STORY PROMPTS

accuracy, mostly reflect the speakers' proficiency levels, rather than the task itself, as supported by findings from Levkina and Gilabert and this narrative task.

Another explanation for the similarity in compositional complexity is that the measures may not have been sophisticated enough to capture any differences in language performance elicited by the different prompts, as mentioned by Levkina & Gilabert (2012). Noun phrases with equal compositional complexity may differ in structural complexity, and it could be said that noun phrases in Race were more complex than in Frog in measures of structural complexity. For instance, a more detailed analysis with measures of structural complexity, reported by Author1 & Coauthor2 (2011), shows that Race elicited more pronouns and complex noun phrases (e.g., with relative clauses), while Frog elicited more coordination (such as *the boy and the dog*). In addition, Race elicited more complementizers (*that* in phrases like *he thinks that*) per AS-unit as compared to Frog. This finding is also consistent with Robinson's predictions about task complexity along the *resource-directing* dimension, in that prompts that require reasoning about characters' beliefs and intentions will encourage the use of stative verbs, such as *think* (Robinson, 2003).

V Conclusion

The findings of the study suggest that even seemingly similar picture prompts can elicit different language performances. While the prompts were selected to be similar in three salient features, there were some unexpected differences in fluency and lexical variety. These findings support a multi-faceted view of narrative task complexity, considering resource-dispersing and

FIVE PICTURE STORY PROMPTS

task difficulty factors in addition to narrative structure, and storyline complexity, and number of elements.

This study illustrates the importance of reporting information about picture story prompts, providing their sources and descriptive details, because specific characteristics of the prompts might unexpectedly affect the language elicited in seemingly equivalent prompts. Such information is crucial not only for evaluating the outcomes of a study properly, but also for conducting replications and follow-up studies. When prompts cannot be selected from previous research, piloting is essential, for one reason, because it can never be assumed that essential words (e.g., *turtle*) or even highly frequent words (e.g., *because* and *can*) can and will be elicited from all participants. In short, researchers as well as language teachers and test developers should not simply assume that seemingly equivalent prompts will elicit similar performance.

References

- Ahmadian, M. J., Abdolrezapour, P., & Ketabi, S. (2012). Task difficulty and self-repair behavior in second language oral production. *International Journal of Applied Linguistics*, 22(3), 310-330.
- Ahmadian, M. J., Tavakoli, M., & Dastjerdi, H. V. (2015). The Combined Effects of Online Planning and Task Structure on Complexity, Accuracy, and Fluency of L2 Speech. *Language Learning Journal*, 43(1), 41-56. doi: 10.1080/09571736.2012.681795
- Bae, J., & Lee, Y-S. (2010). The validation of parallel test forms: "Mountain" and 'beach' picture series for assessment of language skills. *Language Testing*, 28(2), 155-177. doi: 10.1177/0265532210382446
- De Jong, N. & Poelmans, P. (2011, September). *Accuracy and complexity in second language speech: Do specific measures make a difference?* Paper presented at EUROLA Conference, Stockholm, Sweden.
- De Jong, N. & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533-568.
- Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, 19(4), 347-368.
- Ellis, R. (2003). *Task-based Language Learning*. Oxford: Oxford University Press.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. New York: Oxford University Press.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all

FIVE PICTURE STORY PROMPTS

- reasons. *Applied Linguistics*, 21(3), 354-375.
- García Mayo, M. P. (2007). *Investigating Tasks in Formal Language Learning*. Multilingual Matters.
- Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546.
- Heaton, J. B. (1966). *Composition Through Pictures*. Essex: Longman.
- Ishikawa, (2007). The effect of manipulating task complexity along the (±Here-and-Now) dimension on L2 written discourse. In M. P. Garcia Mayo (Ed.), *Investigating Tasks in Formal Language Learning* (136-156). Toronto: Multilingual Matters.
- Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, 62(2), 439-472.
- 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.
- Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In Housen, Kuiken, & Vedder's (Eds), *Dimensions of L2 performance and proficiency*. (171-197). Philadelphia: Jon Benjamins Publishing Company.
- Mayer, M. (1967). *A Boy, a Dog, and a Frog*. New York: Dial Books for Young Readers.
- Mayer, M., & Mayer, M. (1971). *A Boy, a Dog, a Frog, and a Friend*. New York: Puffin Pied Piper.
- McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15(3), 323-337.

FIVE PICTURE STORY PROMPTS

- Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 5-19.
- Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL*, 45(3), 241-259.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555-578.
- Rahimpour, M., & Hosseini, P. (2010). The impact of task complexity on L2 learners' written narratives. *English Language Teaching*, 3(3), 198-205.
- Rescher, N. 1998. *Complexity. A Philosophical Overview*. New Brunswick and London: Transaction Publishers.
- Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, 35(1), 87-92.
- Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45(1). 99-140.
- Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies*, 21(2), 45-105.
- Robinson, P. (2011). 'Second language task complexity, the cognition hypothesis, language learning and performance' in Robinson (Ed.), *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. (3-37). John Benjamins Publishing.
- Rossiter, M. J., Derwing, T. M., & Jones, V. M. (2008). Is a picture worth a thousand words.

FIVE PICTURE STORY PROMPTS

TESOL Quarterly 42, 325-329.

Shafer, D. (2006). *Revolution: Software at the Speed of Thought*. Monterey, CA: Shafer Media.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, lexis. *Applied Linguistics*, 30(4), 1-23.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1): 93-120.

Skehan P., & Foster, P. (2012). Complexity, accuracy, fluency, and lexis in task-based performance: A synthesis of the Ealing research. In Housen, Kuiken, & Vedder (Eds.) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy, and Fluency* (199-220). Philadelphia: John Benjamins.

Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System*, 37(3), 482-495.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 Learners and Native Speakers. *ELT Journal*, 65(1), 71-79.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(s1), 439-473.

Van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press.

Vercellotti, M. L. & Packer, J. (2015). *Shifting structural complexity: The production of clause types in Speeches given by English for academic purposes students*. Manuscript submitted for publication.

FIVE PICTURE STORY PROMPTS

ⁱ A third part of Robinson's model, task condition, concerns interactive factors, which are less relevant for this research with recorded monologues. Interested readers are directed to Robinson (2011).

ⁱⁱ The Michigan Test of English Language Proficiency scores of students placed at this level typically average 60 (*SD* = 11).