

# An Investigation into the Equivalence of Three Pictures for Creative Story Writing: ‘Dog Owners’, ‘Lost Dog’, and ‘Overslept’\*

Heejung Suh

Kyungpook National University

Jungok Bae

Kyungpook National University

Alternate pictures that are proven to be equivalent are in high demand to assess creative thinking and language skills. This study aimed to investigate the equivalence of three pictures (‘Dog owners,’ ‘Lost Dog,’ and ‘Overslept’) recently developed for use in a creative writing task. Middle school students ( $N=183$ ) wrote a story in English based on one of the three prompts distributed randomly. Four writing features (fluency, syntactic complexity, lexical diversity, and temporality) were analyzed with Coh-Metrix and MANCOVA. The three prompts were largely equivalent in their capacity to detect differences among writers in all the features of writing. The difficulty levels of the three prompts, however, were not necessarily the same. Two prompts, Dog Owners and Lost Dog, were verified as equivalent prompts, and therefore, they are recommended as alternate forms to assess creative language skills in repeated measurements. The Overslept prompt had greater facility in eliciting diverse words and more temporal connectives in composing stories. The differential difficulty shown among the prompts suggests that the validity of using different picture versions in repeated assessment remains questionable unless those versions undergo equivalence verification.

**Key Words:** Picture prompts, Creative writing, Stories, Equivalence, Repeated measures

## I. Introduction

Pictures are useful prompts for eliciting creative thoughts and expressions in writing(Bae & Lee, 2011; Crisp & Sweiry, 2006; Shapiro & Hudson, 1991). In assessments to track changes in creative, expressive language skills, it is necessary to use two or more forms of pictures

---

**Corresponding Author:** Jungok Bae(jungokbae@knu.ac.kr)

\*This paper is an expansion and revision of the first author’s Master’s thesis.

that look different but can be used interchangeably and function as equivalent tests. Such forms are known as alternate, equivalent, or parallel forms of a test(American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Once picture forms are validated as equivalent, they can be used as reliable and valid test instruments to make decisions about students' creative writing at multiple time points. To date, however, only a handful of pictures are available that have been verified as equivalent test instruments to assess such skills (cf. Literature Review).

Equivalent forms are defined and established specifically using a few conditions (AERA, APA, & NCME, 1999; Bachman, 1990; McCaffrey, Duff, & Westervelt, 2000): First, they should be created for the same testing purposes and administered using the same testing procedure; Second, equivalence is established when two or more versions yield the same statistical details. Specifically, for alternate test forms to be equivalent, the conventional reliability theory on equivalence (Cronbach, 1947) requires that the means and variances should be equal across the alternate forms. A more thorough investigation of equivalent test forms would require further requirements, such as the equivalence of intercorrelations (or factor structures), but the equivalence of means and variances is the "minimal constraints" on multiple test forms in order for them to qualify as equivalent instruments(McDonald, 1999). This minimal requirement will be tested in the present study.

Based on the principle of equivalence establishment as well as recognizing the demands for equivalent pictures for creative writing assessment, this study aims to:

- (1) Investigate whether the three recently-developed picture prompts are equivalent test forms, that is, whether they satisfy the equality of the means and of the variances;
- (2) Discuss the effect of prompt differences on writing features if the prompts do not meet the equality requirements.

It is commonly known that the mean is a measure of the difficulty (or facility) of a test. The higher the mean, the easier the test. If two or more picture prompts generate equal means, it would mean that the difficulty level of those prompts would be the same. Variance is known as a measure of the spread of the scores from the mean, indicating that variance represents individual differences in test scores. If two or more picture prompts generate equal variances, it would indicate that those prompts would have the same discriminating power.

This study primarily intends to provide validated equivalent picture prompts for assessment researchers and teachers that can be used to track changes in students' writing. However, in any case, if the three prompts of this study fail to be equal in terms of both means and

variances, or one of the two, the study will discuss why the prompts are not equivalent. The study will use such a case to show that students' writing features can be affected by the picture they use. The results will help researchers and practitioners note and address the important issues of equivalence of multiple pictures.

## **II. Literature Review**

### **1. Usefulness of Picture Prompts**

Picture stimuli are useful because they provide writers with a starting point with a topic, a setting, characters, or clues relevant to events and help them organize stories (Berman & Slobin, 1994; Shapiro & Hudson, 1991). When mixed with verbal fragments, picture stimuli can effectively generate language samples (Greenhoot & Semb, 2008).

Numerous studies have compared picture prompts with verbal prompts only, and pictures have been found to be more efficient. For instance, Schneider and Dubé (2005) showed that study participants given picture support produced more fluent stories compared with those without picture support. According to Baker and Quellmalz (1979), picture prompts inspire writers more as a source for their stories than do verbal prompts, and a visual prompt assists in the organization of information. These authors pointed out the limitation of written prompts: Students with poor reading skills may perform poorly because of their inadequate reading comprehension of the topic requirements than because of their actual writing ability. Since testers usually choose pictures that seem interesting to motivate students to perform a required task, pictures can help reduce students' stress and anxiety during test taking. Because tests tend to be stressful for most students, as Crisp and Sweiry (2006) comment, elements that generate their interest or make a test look less daunting, like the use of pictures, could have a beneficial role, particularly in the case of tests designed for less able students.

### **2. Equivalent Forms and Picture Cases**

Equivalent test forms are used in a few assessment situations as illustrated below (McCaffrey et al., 2000; McDonald, 1999). One situation is to track changes in achievement. The same test is given before and after instruction to assess the students' degree of achievement and to decide whether more teaching and learning is required. If testers reuse a test that students have already taken, the problem of students' practice and familiarity (Cliffordson, 2004; Hausknecht, Halpert, Di Paolo, & Gerrard, 2007) and/or memorization (Raymond, Neustel, & Anderson, 2007) is unavoidable. Test scores will likely be better because of the previous exposure to the test and test taking experience, and alternate

test forms that can be used interchangeably are necessary to minimize these effects (Cliffordson, 2004; Duff, 2012; Hausknecht et al., 2007; McCaffrey et al., 2000).

Another occasion is when test takers want to obtain a better score on well-known tests by taking the test more than once. In such cases, the same tests are administered repeatedly, and concern for test security and fairness arises (Oswald, Friede, Schmitt, Kim, & Ramsay, 2005). Therefore, different test forms are necessary for different administration, and it is crucial to verify the equality of these forms for test security and fairness (Weir & Wu, 2006). However, alternate forms are not always available for most testing programs, and many of those available do not meet the equivalence criteria (McCaffrey et al., 2000; Quereshi, 2003).

Of interest to this study, we are dealing with picture prompts that can be used as potential alternate forms of a test. Many studies compared different picture forms and associated performance results. In those studies, the topics depicted in the pictures were either different or similar.

For instance, Pearce (2003) compared a single-scene prompt and a wordless picture book and found that the picture book elicited longer, more informative and complex stories. The topics in the two prompts, however, were different: children looking at cats vs. a frog story.

Shapiro and Hudson (1991) compared two picture booklets: One booklet was a problem-resolution version, and the other with no problem-resolution embedded in the sequence. The author found differences in children's discourse features in response to these booklets.

Similarly, Schweizer (1999) compared a static picture depicting "a delivery man with a box" and a dynamic picture depicting "a cliff rescue." The effect of the picture content was significant; better narratives were generated from the cliff picture than from the box-delivery picture. These studies showed that the variation in the topic or content of the picture was a significant factor in the differences of writing features.

Several other studies compared different pictures with the specific purpose of investigating equivalence across pictures. Pena et al. (2006) tested for the equivalence of two picture books (Bird and His Ring vs. Two Friends) and found that both were equivalent, while the topics, as the picture names indicated, were clearly different.

Bae and Lee (2011) investigated whether two picture series were parallel. One series described "hiking on a mountain," and the other "a picnic at a beach." The two sets were proven to be a parallel series. Bae (2014) also investigated whether two other series, entitled "pizza" and "amusement park" were equivalent forms. These were verified to be equivalent forms. They were equivalent because, in the design stages, the picture series were purposefully designed to be comparable, for instance, in terms of the number of main events and number

of characters. With these elements tightly controlled to be equivalent, the sub-topics of beach vs. mountain, and park vs. amusement park, did not lead to any differences in the students' writing features. Although these topics were superficially different, they all belonged to the larger category of 'a family having a picnic.

The present study is consistent with the studies by Pena et al.(2005), Bae(2014), Bae and Lee(2011), and Weir and Wu(2006) in its explicit purpose to test for equivalence across different picture prompts. However, the current study is different from these studies. While these other studies used a series of pictures, the present study uses single-scene pictures. These single-scene prompts are writing prompts developed for a relatively new task type in which students are asked to compose a story by imagining what happened before, during, and after what the single scene depicted (cf. Method). Furthermore, this study uses middle school students, while most of the studies above used students of younger ages.

Multiple forms of pictures could be developed either with the intention or with no explicit intention to make them equivalent measurement tools, and they could often be used without attending to their real equivalence. While the several studies reviewed above validated equivalent pictures, which can be used as valid tools for longitudinal assessments, more validated pictures are needed to meet the demand for more reliable and valid assessments of writing.

At the same time, if multiple pictures are found to be non-equivalent, researchers may be able to trace possible factors causing the differences. This investigation will help bring about awareness to the issue of non-equivalence and provide suggestions for progress towards equivalent picture development.

### **III. Method**

#### **1. Study Participants**

The participants in this study consisted of second-year middle school students in a public middle school in Daegu, South Korea. They were learning English as a foreign language as a part of the curriculum. This school had the second highest score in the city on the National Assessment of Educational Achievement test as of 2013, the year in which the writing samples were collected. The high performing school was selected because the study needed writing samples of substantial text amount in order to analyze the writing features. Six classes were selected based on the students' English scores on their latest term exam. The six classes had the highest levels of English proficiency at the school. This selection allowed for the collection of writing samples with a sufficient amount of text to analyze.

A total of 242 students participated in the testing. However, writing samples with words less than 100 were excluded from the analysis based on the recommendation by Koizumi and In'nami (2012). These authors demonstrated that some indices such as MTLT, a lexical diversity index used in this study, were not reliable with texts less than 100 words. After this exclusion, the sample consisted of a total of 183 students (133 females and 50 males).




## 2. Writing Prompts

**Picture Prompts Developed.** Three picture prompts (Figure 1) were used in a creative story-writing task in this study: Dog Owners, Lost Dog, and Overslept. The genre of the task is story writing, and the story genre has been supported because it promotes discourse skills and creative personality in children(Lee & Lee, 2008; Peterson & Dodsworth, 1991).

Each picture had a single-scene intended to evoke a story. In the Dog Owners picture, two persons were walking their dog in the park, and one dog fell in love with the other dog, going after it, which embarrassed its owner. In the Lost Dog scene, a boy holding a poster of a lost dog was asking a grandmother sitting on a bench eating snacks. In the Overslept event, a boy was oversleeping, and a mother was looking at him holding a breakfast tray.

As shown in the written directions, the task required the writers to compose a story in English, describing the current event depicted in the single scene and imagining what happened before, now, and after the event. The prototype task with these directions for a single-scene picture was first developed in 2009 for use in an entrance test for a university-housed gifted program to select students skilled in verbal creativity: the process of developing this prototype task is detailed in Bae, Jordahl, and Lee(2012).

The three pictures in Figure 2 were developed based on the attributes of the prototype task mentioned above. Dog Owner and Overslept had been developed as a part of a project in an undergraduate course entitled "Assessment Theory in English Education" taught by the corresponding author in 2012. The two pictures were among a dozen pictures developed according to the specifications authored by the corresponding author who was one of the joint developers of the afore-mentioned prototype task. Lost Dog was developed in 2011 for use in a writing assessment for children enrolled in an English program at a university. These three pictures look different, but they all were intentionally designed to be equivalent in terms of the format, number of characters, and presence of a stimulating incidence, following the prompt characteristics described in the task specifications.

<p><u>Directions</u><sup>§</sup> (common for all pictures):</p> <p>“Make up a story around the pictured scene in English. What happened before this? What is happening now, and why? What will happen next? Use your imagination! Be as creative as you can! Have fun with this! However, your story should go with the picture.”</p>	 <p style="text-align: center;"><b>Dog Owners</b></p>
 <p style="text-align: center;"><b>Lost Dog</b></p>	 <p style="text-align: center;"><b>Overslept</b></p>

<sup>§</sup>Directions: Similar verbal directions also appear in prior studies that used the same prototype task but with different pictures, such as Jung & Bae(2013), Bae, Bentler, & Lee(2016), and Bae et al.(2012).

<sup>§§</sup>Picture developers: Dog Owners: Jung-Wuk Kim & So-Hye Shon; Lost Dog: Jungok Bae & Jonathan Jordahl; Overslept: Joon-Ho Lee and Ga-Hyun Park (Content) & Hyo-Jung Jung (Illustrations). Used and reprinted with permission from the picture developers.

[Figure 1] Picture prompts used in this study<sup>§§</sup>

Dog Owners and Overslept were selected in the current equivalence investigation because they were found popular among students from a survey on prompt preferences. Lost Dog was included in this study because the prompt turned out to be useful for eliciting writing samples from previous test administrations. The current study wanted to test for the equivalence of these three prompts for the previously specified research purposes.

### 3. Test Administrations

The writing test was administered in 2013. The test materials consisted of the writing prompts with a writing pad. Written guidelines on test administration were provided for

several English teachers who were in charge of English teaching for the six classes participated in this study. The guidelines were provided before the day of the testing to help the teachers understand the testing procedures and ensure consistent test administrations across the different classrooms.

The three writing prompts were randomly distributed to the students in each classroom. To facilitate this process, the pictures and instructions were printed in three distinguishable colors: white (Dog Owners), pink (Lost Dog), and green (Overslept). For all six classrooms, the English teacher in each classroom distributed the colored papers with a rotation of white, pink, and green from the first student to the last one seated in rows. This distribution divided the students into three picture groups each of which received a different picture. Through randomization, each student had the equal probability (DeCoster, 2006) of receiving any of the three pictures. While randomization does not guarantee that the groups will have the same important characteristics, it is still a good way of equating the groups(DeCoster, 2006). Through the random assignments, the groups were distinguished by the picture they received, but were largely comparable in terms of other confounding variables (such as ability, gender, socioeconomic status, and so forth).

Approximately 15 minutes were allotted for the picture distribution and a short explanation about the test by the teachers, following which 30 minutes were allowed for writing. Instructions on the paper were read aloud by the teachers and the students in both English and Korean during the first 15 minutes to clarify what the students should do on the task.

#### 4. Variables

The independent variable was the picture group with the three picture prompts (Dog Owners, Lost Dog, and Overslept). Four writing features were selected as the dependent variables: fluency, syntactic complexity, lexical diversity, and temporal connectives, defined below.

Fluency was included in the analysis as an essential writing feature. Wolfe-Quintero, Inagaki, and Kim(1998) defined fluency as the total number of words in a text. If the same task is used and the same time is allotted for all writers, the amount of writing is one indicator of writing proficiency (Bae, 2007). More fluent writers may produce a greater number of words. Thus, word count, was used as an indicator of fluency.

Syntactic complexity was included as the indicator of grammatical quality of writing. Studies (e.g., Lee & Koh, 2013) have found that grammar and vocabulary had a significant effect on gifted students' language ability. It was appropriate to include syntactic complexity, along with lexical diversity (below), in the analysis of creative writing in this study. As an indicator of syntactic complexity, 'the number of words before the main verb' was used based



on McNamara, Crossley, and McCarthy (2010), who showed that this index best represented syntactic complexity.

Lexical diversity (LD), or vocabulary diversity, refers to the range of different words used by a writer (McCarthy & Jarvis, 2007, 2010). Many studies attest that LD is significantly associated with writing quality (e.g., Laufer & Nation, 1995). There are several indicators of LD, such as TTR, vocd, and MTL. Among them, MTL was selected in this study, because MTL is least-affected by text length whereas other LD indices are susceptible to text length (McCarthy & Jarvis, 2010).

Temporal connectives refer to time-indicating conjunctions, adverbs, and adverbial phrase such as before, and then, until, later, soon, and after a month (Department for Education and Employment, 1998; Halliday & Hasan, 1976; Kim, Koo, & Bae, 2015). They link parts of language, establishing temporality, time-associated conditions. The temporal dimension is important in stories because stories consist of sequenced events. In the sequence of events, event cohesion is established using a variety of temporal connectives (Duran, McCarthy, Graesser, & McNamara, 2007). Therefore, temporal connectives were included as a measure of event cohesion.

## 5. Scoring and Analysis

To measure the four writing features, this study used the Coh-Metrix. The Coh-Metrix program analyzes texts on various linguistic and discourse features (Graesser, McNamara, Louwerse, & Cai, 2004). For analysis with the Coh-Metrix, the handwritten stories were typed into MS word files. During the typing procedure, misspellings were corrected but not grammatical errors. The converted texts were analyzed with the Coh-Metrix for the four writing features.

SPSS was used for data analysis. To determine the similarities or differences in the four writing features among the three picture groups, multivariate analysis of variance with a covariate (MANCOVA) was used. As a covariate, the students' scores from the latest mid-term exam were used: these scores represented the students' existing writing ability to be controlled for, that is, to be statistically held constant across the three groups. This study controlled confounding variables through the random assignment of the pictures mentioned previously, and the use of the covariate was an additional attempt to control the primary confounding variable, students' existing English writing ability around the time of the main task, the story-writing.

## IV. Results

### 1. Descriptive Statistics and Correlations

Table 1 presents the descriptive statistics for the four writing features categorized into the three picture groups. As indicated in Table 1, skewness and kurtosis were approximately within the range of +/- 2. Therefore, they satisfied the normality assumption (Lomax, 2001) for using correlations and MANCOVA.

<Table 1> Descriptive Statistics (N=183)\*

Variables	Group by Picture Use	Mean	SD	Kurtosis	Skewness
Fluency (word count)	Dog Owner	221.79	71.34	-.69	.39
	Lost Dog	243.32	85.48	-1.06	.18
	Overslept	224.86	78.38	.41	.55
Syntactic complexity	Dog Owner	2.23	.62	2.07	.95
	Lost Dog	2.02	.64	.55	.89
	Overslept	2.20	.83	.04	.67
Lexical diversity	Dog Owner	46.40	15.36	2.19	.74
	Lost Dog	47.11	12.08	-.43	.19
	Overslept	56.50	15.83	2.30	1.14
Temporal connectives	Dog Owner	35.97	14.91	.57	.44
	Lost Dog	38.80	17.69	2.20	1.17
	Overslept	53.16	19.41	-.19	.22

\*N=56 (Dog Owners), 68 (Lost Dog), 59 (Overslept)

Table 2 presents the extent to which the variables of this study, each of the four writing features and the covariate, were related to one another. The intercorrelations were relatively weak; nevertheless, Bartlett’s test of sphericity rejected the hypothesis that all the correlations in the correlation matrix were zeros (approximate chi-square=41.80,  $p<.001$ ). Therefore, this result met another assumption for using MANCOVA—that there are significant, if not moderate, correlations among the dependent variables (Meyers, Gamst, & Guarino, 2006). In addition, the covariate, the scores on the latest term exam, had significant relationships with some of the dependent variables, indicating that the covariate influenced those variables, and therefore, was appropriate for use to control for its probable effects on those writing features.

<Table 2> Pearson Correlation Matrix (N=183)

	Fluency (Word count)	Syntactic complexity	Lexical diversity	Temporal connectives
Syntactic complexity	.012			
Lexical diversity	.271**	-.103		
Temporal connectives	-.099	.164*	.254**	
Covariate	.297**	.064	.372**	.116

\*\*Significant at alpha=.001, \*Significant at alpha=.05.

## 2. Equivalence of Means

This section addresses whether the three groups were equivalent in terms of means. This equality was examined by comparing the means for the four writing features among the three groups, which was the main function of the MANCOVA procedure used. The results of the MANCOVA multivariate test are provided in Table 3.

<Table 3> Multivariate test

Effect		value	<i>F</i>	hypothesis <i>df</i>	error <i>df</i>	<i>Sig.</i>
Covariate	Pillai's	.193	10.49	4	176	.000
	Wilks'	.807	10.49	4	176	.000
	Hotelling's	.238	10.49	4	176	.000
	Roy's	.238	10.49	4	176	.000
Picture Groups	Pillai's	.246	6.20	8	354	.000
	Wilks'	.763	6.37	8	352	.000
	Hotelling's	.299	6.54	8	350	.000
	Roy's	.254	11.25	4	177	.000

The effect of the covariate on the four writing features was significant,  $F(4, 176)=10.49$ ,  $p<.001$ , as reported by Wilks' statistics, the most commonly reported multivariate statistic (Liu, 2002). This was an expected result because the covariate had significant correlations with certain dependent variables (Table 2). The result confirmed that the effect of the covariate was appropriate to partial out. After eliminating the effect of the covariate, i.e., holding the groups' mid-term English scores constant, the main effect of picture variations on the means for the four writing features, taken together, was significant,  $F(8, 352)=6.37$ ,  $p<.001$ , Wilks' statistic. As a result, the null hypothesis that the three picture groups had equal means for the four writing features was rejected. Not all the means were the same.

To find out where the significant mean differences lay, post-hoc comparisons were performed (Table 4). In Table 4, with respect to fluency and syntactic complexity, all three picture groups had statistically similar means. However, as for lexical diversity, the Overslept group had a significantly higher mean than those of the other two groups, which had no difference in their means. The same result was found for temporal connectives.

<Table 4> Post-hoc Comparisons

Variables	Picture group (I)	Picture group (J)	Mean difference (I-J)	<i>p</i>
Fluency	Dog Owners	Lost Dog	-24.88	.069
	Dog Owners	Overslept	-1.35	.924
	Lost Dog	Overslept	23.53	.082

Syntactic complexity	Dog Owners	Lost Dog	.205	.110
	Dog Owners	Overslept	.027	.837
	Lost Dog	Overslept	-.177	.160
Lexical diversity	Dog Owners	Lost Dog	-1.43	.556
	Dog Owners	Overslept	-9.79*	.000
	Lost Dog	Overslept	-8.29*	.001
Temporal connectives	Dog Owners	Lost Dog	-3.08	.330
	Dog Owners	Overslept	-17.07*	.000
	Lost Dog	Overslept	-13.99*	.000

\*The mean difference is significant at alpha=.05

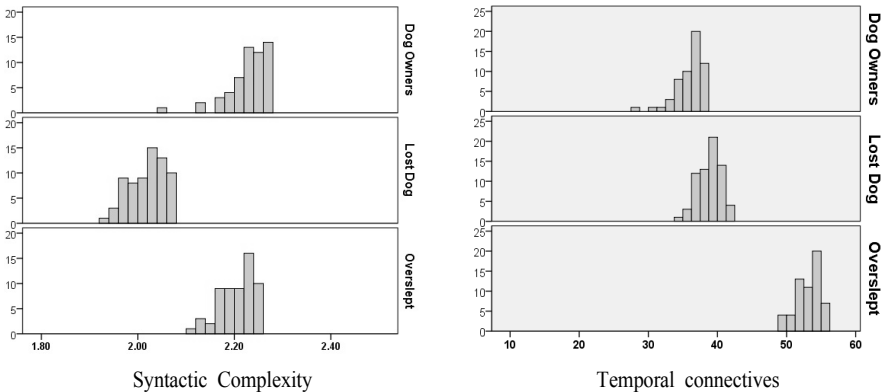
### 3. Equivalence of Variances

This section addresses whether the three picture prompts generated the same variances. For this purpose Levene’s statistics within the MANCOVA procedure were inspected (Table 5). In addition, the graphs in Figure 2 show how widely spread out the scores were. For these graphs, the adjusted individual scores, obtained after controlling for the students’ covariate scores, were used. Only two variables, syntactic complexity and temporal connectives, were selected for this illustration due to space constraints.

<Table 5> Levene’s Test of Equal Variances

Features	F	p
Fluency (word count)	.857	.426
Syntactic complexity	3.116	.047
Lexical diversity	1.621	.201
Temporal connectives	1.908	.151

df1=2, df2=180, all variables



[Figure 2] Similar degrees of score variations across three picture groups

Looking at the results of Levene's test of equal variances (Table 5), all  $p$  values were well beyond .05, the criterion for nonsignificant differences in groups: An exception was the  $p$  of .047 for syntactic complexity, which fell a little short of .05. If we look at the standard deviations (Table 1), the scores for Dog Owner and Lost Dog were almost equally spread out (.62, .64, respectively), but the spread of the scores for Overslept was slightly greater (.83). By the  $p$  of .047 and the differences in the SDs, the Overslept group would be considered different from the other groups. However, these decimal differences seem minor if we look further at the distributions of the syntactic complexity scores in Figure 2; the ranges of the scores were largely the same across all three picture groups even for syntactic complexity; it was hard to say that the range of the scores for syntactic complexity for the Overslept group was any larger than those for the other picture groups. It is thus concluded that overall, the range of the scores generated by each prompt was more or less similar for largely all the four writing features.

## V. Conclusion and Discussion

The primary purpose of this study was to verify the equivalence of the three single-scene pictures (Dog Owners, Lost Dog, and Overslept), which were developed as homogeneous pictures following a fairly new type of story-writing task, called 'Imagine Before, Now, and After' the event depicted (Bae et al., 2012; Jung & Bae, 2013). Taken all together, the findings are summarized as follows, and their interpretations follow.

### 1. Largely Equivalent Variances and Partially Equivalent Means

This section concludes Research Question 1, which asked about the equality of means and variances required for equivalent test forms, and interprets the results.

With respect to the variance equality, the three picture prompts generated more or less similar variances in the scores for the four writing features (fluency, syntactic complexity, lexical diversity, and temporal connectives). This result indicates that the three prompts had similar capacities to detect differences among individual test takers in these writing features; all three prompts had similar discriminating power.

However, with respect to the mean equality, all the prompts did not generate statistically similar means, indicating that the difficulty (or facility) of the three prompts was not necessarily equivalent depending on which writing features were examined. The three prompts yielded the similar means for fluency and syntactic complexity. However, they had a different mean for lexical diversity and temporal connectives: The Overslept prompt had the highest mean for both lexical diversity and temporal connectives, and the other two prompts had

equally lower means for both writing features. This result indicates that the writers who were assigned the Overslept prompt produced more elaborately written text with diverse words and clearer descriptions of event sequences by using a greater number of time-associated connectives. The Overslept picture was therefore more facilitative in the writers' use of the diverse words and temporal connectives.

In conclusion, the findings are summarized:

- (1) Discriminating power: The three prompts were similarly capable of detecting individual differences among the writers;
- (2) Prompt difficulty: Not all three prompts were equally difficult for the writers creating a story; the Overslept prompt was more facilitative in eliciting diverse vocabulary and establishing temporal cohesion than were the other prompts;
- (3) The prompts found equivalent: The other two prompts, Dog Owners and Lost Dog, nonetheless, were verified to be equivalent prompts for all the writing features examined in this study. While these prompts were less facilitative in eliciting diverse words and making the stories temporally cohesive, they nevertheless had the same levels of prompt difficulty and the same discriminating powers. Therefore, Dog Owners and Lost Dog are safely recommended for use as alternate prompts in future repeated measurements.

In the design stage, except for the primary variable of picture differences, other confounding variables were largely controlled for with the covariate scores and the random assignment of the pictures, as well as the same testing procedure. Therefore, it is safe to say that the above findings are more or less strong.

## 2. Students' Topic Familiarity Affecting Writing Features

This section addresses the second research aim—discussing the prompt differences affecting different performance results as partly concluded in the previous section. The following question may arise: What aspects of the picture difference might have influenced the prompt difficulty, or facility?

The reason for why the Overslept prompt elicited more diverse words and more temporal connectives may have to do with topic familiarity. The picture, Dog Owner, depicts a story about one male dog going after a female one; while quite interesting, stories relating to dogs falling in love are not readily found and therefore less familiar to the students.

Another picture, Lost Dog, while also interesting, was not experienced by everyone. A familiar task tends to free up learners' cognitive resources and facilitate planning, decreasing

task difficulty; in contrast, an unfamiliar topic and time pressure raise communicative stress, increasing the cognitive complexity and making the task more difficult (Skehan, 1996).

The other prompt, Overslept, however, depicts a frequent topical experience known to everyone. When writers compose, they use their own experience, knowledge, and understanding with respect to a particular topic to create meaning (Cleaver, Scheurer, & Shorey, 1993; Kintsch, 2005). The Overslept prompt definitely depicts a topic that most people encounter frequently in daily life. When the writers first received the picture, they could have easily related to the scene. Developing writers can list everything they already know about a topic (McCutchen, 1996), and writers use past experience, knowledge, and understanding to extend the narratives of the picture (Cleaver et al., 1993). Their knowledge on the topic of oversleeping might have facilitated the production of related ideas which led to the use of more diverse words in their writing samples (Ackerman, 1991; McCutchen, 1986; Young & Leinhardt, 1998). The familiarity with the topic also facilitated the coherent organization of the events in time sequence, which in turn, facilitated the use of time-indicating conjunctions and adverbials.

The highest result in lexical diversity and temporal cohesion for the Overslept prompt can further be explained with frame semantics (Fillmore, 1982). A frame (a general cover term similar to schema) is "any system of concepts related in such a way that to understand any one of them, you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available" (Fillmore, 1982, p. 111). The scenes represent the categories of experiences. The writers' frequent personal experiences with oversleeping may have provided the conceptual frame. Once the writers established a frame prompted by the scene, all other ideas in the frame were made available in the writers' mind, generating diverse ideas as well as facilitating the narration of events using a variety of temporal connectives.

In conclusion, topic difference in the pictures is considered a possible reason for the non-equivalent difficulty of the pictures in eliciting certain writing features such as temporal connectives and lexically diverse words.

### 3. Future Prospects

One limitation of the present study needs to be mentioned. This study did not assess content quality in the stories. Content is an important quality of writing, and by scoring content, aspects such as creativity, interestingness, and relevance, which are part of content attributes (Bae, 2007; Bae et al., 2016), could have been incorporated in the analysis. Future cross-validation could use human scoring of content quality in the equivalence investigations.

Despite the limitation, the present study has made a few contributions. First, this study has

provided two equivalent picture prompts, namely Dog Owners and Lost Dog, for future studies to use in their repeated assessments of expressive, creative language skills. If used, the prompts will help enhance the validity of those assessments.

The second contribution of the present study is that the study results draw attention to the issue of picture differences, reconfirming the findings of prior studies reviewed—topic variation in picture prompts can have a significant effect on story performance(Cleaver et al., 1993; Cykowicz, Friedman, Snodgrass, & Rothstein, 1994; Pearce, 2003; Schweizer, 1999). This study has indicated that future developers and users of equivalent pictures should consider writers' familiarity with the topic in the pictures as an important factor in writing features.

Third, the findings of this study help us understand the need for actually testing the equivalence of different picture prompts. This study showed that different picture prompts, although developed to the same task-specifications, can, in fact, turn out not to be equivalent. Therefore, the findings provide a warning that alternate picture prompts must undergo a validation to ensure their equivalence before use and to prevent any unfair use of and decisions about scores based on such prompts. With these contributions presented above, the present study hopes to ultimately contribute to more reliable and valid writing assessments that use pictures to compare students' language samples at multiple time points.

## References

- Ackerman, J. M. (1991). Reading, writing, and knowing: The role of disciplinary knowledge in comprehension and composing. *Research in the Teaching of English*, 25(2), 133-178.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.
- Bae, J. (2007). Development of English skills need not suffer as a result of immersion: Grades 1 and 2 writing assessment in a Korean/English Two-Way Immersion Program. *Language Learning*, 57(2), 299-332.
- Bae, J. (2014). The confirmation of equivalence of two picture series for children's story writing: 'Pizza' and 'Amusement Park.' *Journal of Educational Evaluation*, 27(1), 209-229.
- Bae, J., & Lee, Y. S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing*, 28(2), 155-177.
- Bae, J., Bentler, P. M., & Lee, Y. S. (2016). On the role of content in writing assessment. *Language*



- Assessment Quarterly*, 13(4), 302-328.
- Bae, J., Jordahl, J., & Lee, Y. S. (2012, May). *The effect of relative simplicity or complexity of picture prompt on language and creativity performances in writing*. Paper presented at the 2012 Annual International Conference of the Korea English Language Testing Association, Seoul, Korea.
- Baker, E. L., & Quellmalz, E. (1979). *Results of pilot studies: Effects of variations in writing task stimuli on the analysis of student writing performance. Studies in measurement and methodology. Work unit 1: Design and use of tests*. Los Angeles: University of California at Los Angeles.
- Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Erlbaum.
- Cleaver, B., Scheurer, P., & Shorey, M. (1993). *Children's responses to silhouette illustrations in picture books*. Rochester NY: International Visual Literacy Association.
- Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment*, 20(3), 192-204.
- Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, 48(2), 139-154.
- Cronbach, L. J. (1947). Test 'reliability': Its meaning and determination. *Psychometrika*, 12(1), 1-16.
- Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology*, 65(2), 171-237.
- DeCoster, J. (2006). *Testing group difference using T-tests, ANOVA, and nonparametric measures*. Retrieved June 30, 2016 from <http://www.stat-help.com/notes.html>.
- Department for Education and Employment (1998). *The National Literacy Strategy: a framework for teaching*. London: Author.
- Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, 27(3), 248-261.
- Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, 39(2), 212-223.
- Fillmore, C. (1982). Frame semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the morning calm* (pp. 111-137). Seoul: Hanshin.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.
- Greenhoot, A. D., & Semb, P. A. (2008). Do illustrations enhance preschoolers' memories for stories? Age-related change in the picture facilitation effect. *Journal of Experimental Child Psychology*, 99(4), 271-287.

- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373-385.
- Jung, J., & Bae, J. (2013). The influence of picture prompt variation on writing performance: ‘Series’ vs. ‘Imagine Before and After.’ *English Language Teaching, 25*(2), 27-46.
- Kim, H., Koo, S., & Bae, J. (2015). Positive, negative, and nil effects of connectives in written stories: Analysis by proficiency groups. *Linguistic Research, 32*, 105-124.
- Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders, 25*(1), 51-64.
- Koizumi, R., & In’nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*(4), 554-564.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.
- Lee, S-D., & Koh, W-J. (2013). The structural model analysis of related variables on English reading comprehension ability of gifted students. *The Journal of the Korean Society for the Gifted and Talented, 12*(1), 53-80.
- Lee, K., & Lee, Y. (2008). The effects of language activity programs using fairy tales on progress in children’s creativity. *The Journal of the Koran Society for the Gifted and Talented, 7*(1), 29-46.
- Liu, Y. (2002). Analyzing RM ANOVA related data using SPSS 10. *Measurement in Physical and Exercise Science, 6*(1), 43-60.
- Lomax, R. G. (2001). *Statistical concepts: A second course for education and the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- McCaffrey, R. J., Duff, K., & Westervelt, H. J. (Eds.). (2000). *Practitioner’s guide to evaluating change with neuropsychological assessment instruments*. New York: Kluwer Academic/Plenum Publishers.
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488.
- McCarthy, P. M., & Jarvis, S. (2010). MTL-D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods, 42*(2), 381-392.
- McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*(4), 431-444.
- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*(3), 299-325.
- McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a Theory of Writing. *Educational Psychologist, 35*(1), 13-23.

- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.
- Meyers, L.S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: design and interpretation*. London: Sage Publications.
- Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods, 8*(2), 149-164.
- Pearce, W. M. (2003). Does the choice of stimulus affect the complexity of children's oral narratives? *Advances in Speech-Language Pathology, 5*(2), 95-103.
- Pena, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*(5), 1037-1057.
- Peterson, C., & Dodsworth, P. (1991). A longitudinal analysis of young children's cohesion and noun specification in narratives. *Journal of Child Language, 18*(2), 397-415.
- Quereshi, M. Y. (2003). Absence of parallel forms for the traditional individual intelligence tests. *Current Psychology, 22*(2), 149-154.
- Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*(2), 367-396.
- Schneider, P., & Dubé, R. V. (2005). Story presentation effects on children's retell content. *American Journal of Speech-Language Pathology, 14*(1), 52-60.
- Schweizer, M. L. (1999). *The effect of content, style, and color of picture prompts on narrative writing: An analysis of fifth and Eighth grade students' writing*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.
- Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology, 27*(6), 960-974.
- Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.
- Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: Case study of a semi-direct speaking test. *Language Testing, 23*(2), 167-197.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii at Manoa.
- Young, K. M., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written Communication, 15*(1), 25-68.

## Appendix: Selected Best Sample

Jenny is missing!

ID: ○○○

3 months ago, Tom's dog 'Jenny' ran away. While Tom opened the door to greet his friend Harry, Jenny ran away. Tom tried to find her but he couldn't. He asked every people in his town, but no one knows about Jenny. Tom was so disappointed. After 2 months, he gave up to find his dog Jenny, and he almost forgot about her. However, he saw something strange in the morning, while he going to the school. He saw a poster which is looking for a dog. He just walked away without any thoughts, but he felt strange. Tom thought that the dog's photo in the poster is familiar. After the school, he came back home and walked the same street. He found a wall, on which the poster is attached. Surprisingly, the dog's photo in the poster is same with Jenny! He detached the poster and came home. His family were surprised too. Everyone said "Isn't it Jenny?" Tom was excited. Maybe he could find her! Tom saw a phone number which is written on the poster. He called, and a woman said, "Hello? This is Sara." Tom told to the woman "I want to meet you, and maybe I could tell about the dog you are looking for." So, the woman and Tom promised to meet in the park. Tom took the poster to the park. In the park, woman was waiting for Tom, eating cola and French fries. Tom show the poster and told her "Your dog's photo is exactly same with my dog." The woman, Sara said, "Really? I saw her in the park, so I was taking care of her." "When did you find her? Wasn't it 3 months ago?" Yes, I think so." "And didn't she have a brown spot in her neck?" "That's right!" Tom was excited that he thought he could find her. "But she was gone last night, and I was so worried." Then, Sara's phone rang. "Hello? I found your dog." Tom and Sara ran to the man, who called her. He pointed the dog, and it was Jenny! Jenny recognized Tom and Sara, and she barked and shaked his tail happily. What happened next? Tom's family invited the man and Sara to dinner, and they played happily with Jenny. Until now, Sara sometimes visits Tom's house and take care of Jenny.

= Abstract =

## 창의적 이야기 작문용 세 그림의 동형 조사: 'Dog Owners,' 'Lost Dog,' 'Overslept'

서 희 정

경북대학교

배 정 옥

경북대학교

창의적 사고와 언어기술을 평가하는데 동형검사로 판명된 대체 그림들이 절실히 요구되고 있다. 본 연구는 창의적 쓰기 과제용으로 최근 개발된 세 그림(이름: 'Dog Owners,' 'Lost Dog,' 'Overslept')이 동형 검사지가 되는지 조사하였다. 183명의 중학생들이 무작위로 배분된 세 그림 중 하나에 의거하여 영어로 이야기를 작성하였다. 작문은 네 가지 쓰기요소(유창성, 어휘 다양성, 구조 복잡성, 그리고 시간성)에 대해 Coh-Metrix와 MANCOVA로 분석되었다. 이 세 그림은 변별력에 있어 대체로 위 모든 요소에 대해 비슷하였다. 그러나 이들의 난이도는 요소별로 볼 때 반드시 같지는 않았다. Dog Owners와 Lost Dog 그림은 변별력과 난이도에 있어 동형으로 판명되었다. 그러므로 이 두 그림은 반복 측정에서 타당한 동형 검사지로 추천된다. Overslept 그림은 다양한 어휘와 시간 연결사들을 유발시키는 데에 다른 두 그림 보다 용이하였다. 그림의 난이도가 다를 수 있다는 결과는 반복시험에서 대체 그림을 사용할 시 이들 그림이 동형 검정을 거치지 않고서는 그 타당성이 의심스러울 수 있음을 환기시켜 준다.

주제어: 그림 촉진제, 창의적 쓰기, 이야기, 동형, 반복 시험

1차 원고접수: 2016년 11월 16일
수정원고접수: 2016년 12월 20일
최종게재결정: 2016년 12월 27일