Web-Based Creativity Assessment System that Collects Both Verbal and Figural Responses: Its Problems and Potentials

Jiajun Guo

***Abstract*—Creativity tests have been administered in traditional paper and pencil format for more than a half century. With the prevalence of computer/web-based testing and increasing demands for large-scale, faster, and more flexible testing procedures, it is necessary to explore and test the usability of web-based creativity tests. Yet few studies have focused on the use of technologies in the assessment of creativity. The purpose of the present study was to design and test the feasibility of an online creativity assessment system that can collect both verbal and drawing responses. The following two aims were addressed: (1) evaluate reliability evidence of creativity test scores, and (2) compare the online test with its paper version regarding creativity scores. One hundred and sixty-four participants were recruited from a northeast university in the US and randomly assigned into three groups: online-basic (OB), online-advanced (OA), and paper-and-pencil (PP). The findings indicated that no differences were found between different modes (online vs. paper) or different interfaces (simple tools vs. advanced tools) in terms of either creativity scores or reliability evidence. Additionally, males were found to produce overall significantly higher originality scores than females did in the line meaning test and the real-world problem test. The implications of these findings are further discussed in the paper.**

***Index Terms*—Computer tests, creativity, drawing, visual imagination.**

I. Introduction

With the world becoming increasingly complicated and complex, creativity has been recognized by many as the most important quality for success in the 21st century workplace [1]–[5]. At the same time, our current era is also marked by exponential growth in technological innovation. Although recent years have witnessed the impact of technology on creativity teaching practices in classrooms, few studies have looked at the use of technology in the assessment of creativity. With the development of digital technologies, many standardized tests can now be administered online or on computers, including TOEFL, GRE, SAT, Smarter Balanced Assessments, and even intelligence tests. However, most assessments of creativity are still administered in the same way as they were 50-60 years ago, in the form of traditional paper and pencil tests, making it difficult for them to serve new and emerging purposes like large-scale testing, flexible scheduling, and faster data collection.

Taking the divergent thinking task, a popular assessment of creativity, as an example, the test has been administered on paper since its emergence in 1960s, with only a few attempts to computerize it in recent years. Some researchers have chosen to devise an application or program specifically for the DT tasks, such as Kwon's (1996) *computerized Torrance Tests of Creative Thinking* (TTCT) figural forms [6], Pretz and Link's (2008) *Creative Task Creator* (CTC) [7], Cheung and Lau's (2010) *electronic Wallach-Kogan Creativity Tests* (e-WKCT) [8], Palaniappan's (2012) *Creativity Assessment System* (CAS) [9], and most recently, Zabramski's (2014) computerized-multi-input TTCT figural forms [10]. Creative Task Creator (CTC) is a Java-based program that can generate HTMLs through which participants can complete an alternative uses task. The program can also collect data from different sites and store them on a server for researchers to manage. Similarly, the Creativity Assessment System (CAS) is a web application developed on the basis of the Torrance Test of Creative Thinking (TTCT). CAS is able to automatically calculate fluency, originality, and flexibility scores. Interestingly, an elaboration score was intentionally left out due to issues with the complex programming required. Another computer program for the divergent thinking test was designed by Cheung and Lau, who computerized all the Wallach-Kogan Creativity Tests (WKCT) and administered them in computer labs. The authors found no differences in evidence of reliability between the results from the electronic and paper versions of the WKCT. The study was also the first to provide reliability evidence for the comparability of the two versions of creativity tasks.

As building a computer application is often time-consuming and expensive, other researchers have turned to alternative ways to put creativity tests on computers. With the recent development of online survey services, administering computer-based creativity tests has become much more convenient. For example, Pásztor, Molnár, and Csapó (2015) were able to collect nearly 2,000 responses from 97 classes at 78 schools with the help of an online survey system [11]. Hass (2015) designed and delivered his survey-like creativity tests to online participants using the Qualtrics system so that the testing data could be collected through the internet [12]. He also compared scores obtained from online participants with those obtained in person and found that online participants gave fewer responses than participants taking the test in person, a different result from that of Lau and Cheung (2010).

The difference between online and in-person creativity test results found in Hass's (2015) study indicate the possibility of a mode effect on test scores. Although few studies have

Manuscript received June 6, 2018; revised July 25, 2018.

J. Guo is with the Faculty of Education, East China Normal University, China (e-mail: jjguo@dedu.ecnu.edu.cn).

doi: 10.18178/ijiet.2019.9.1.1168

looked at the mode effect in creativity testing, the mode effect in other types of performance testing has been studied by researchers since the inception of computerized testing. The previous literature has identified several factors that may be responsible for differences in some performance tests, including participant factors (such as computer familiarity and gender) and technological factors (such as screen size and resolution, and item presentation) [13].

What makes things more complicated is that mixed results have been found among the few studies that included computerized figural forms of the creativity test. Kwon's (1996) study was among the first efforts to put the paper creativity test onto computers. Using Hypercard, a programming tool provided by Apple Inc., Kwon was able to develop a computer system for the creativity test and compare the results between paper figural forms and computerized figural forms. He found significant differences between the two modes. Zabramski and his colleagues, on the other hand, found no significant differences in creativity scores between various kinds of input modes, including paper-and-pencil, stylus, mouse, and touch-input [14], [15].

According to Leeson's (2006) framework, there are at least two reasons for the inconsistencies between Zabramski's (2014) and Kwon's (1996) studies. The first reason, which is technological in nature, concerns the availability of relevant tools for drawing tasks. Even for verbal tasks, no consensus exists regarding which technology tool should be used and how it should be used. For drawing tasks, various kinds of input devices (such as mouse, hand-touch, and stylus) and software (such as computer applications and browser-based interfaces) can be used to collect responses. The differences in tools might contribute to the differences in creativity scores between the two studies. The second reason, which is more related to participant characteristics, concerns the confounding influences of participants' technology skills and experience. For example, drawing requires fine motor skills, and the motor skills applied in hand drawing and computer drawing might be different. It is possible that people who often use computers to draw or edit visual contents may possess better skills due to frequent practice, thus making them more fluent in creating computer drawings. Other technology experience not related to drawing may also influence creativity performance, such as playing video games [16]. All of the abovementioned issues may affect the psychometric properties of an online test, including evidence of reliability and validity.

Taking the above issues into consideration, the present study aims to test the feasibility of an online creativity test protocol, and to investigate the psychometric properties of this online creativity test compared to the paper-and-pencil version. Specifically, we will address two research question, including question 1 (*How reliable are the scores obtained from the creativity tests?*) and question 2 (*How do the results obtained from an online creativity test differ from the results of the paper-and-pencil version?*). On the basis of these research questions and the previous literature, two general hypotheses were proposed, including hypothesis 1 (An online creativity test can produce the same reliable creativity scores as the paper test.) and hypothesis 2 (Results obtained in an online creativity test may be different from the results from the paper-and-pencil version.).

II. Methods

A. Design and Participants

A major purpose of this study was to investigate whether there are any differences in terms of psychometric properties between paper-and-pencil and computerized versions of the creativity assessment tool. Therefore, a three-group, between-subjects experimental design was used in the present study, which involved a paper-and-pencil group (PP), an online basic group (OB) and an online advanced group (OA).

The present study was carried out at a northeastern university in the US. Participants were recruited through university public announcement systems such as Daily Digest, Student News, and listserv. A survey link was distributed via these systems to recruit potential participants (18 years of age or older) to sign up for the study. The link directed them to an interface where an information sheet/consent form was presented.

One hundred and sixty-four responses were collected in the current study. Of these respondents, 109 were females, and 55 were males. In addition, 72 were undergraduate students, 85 were graduate students, and 7 were college staff or faculty members.

B. Instruments

To answer the question regarding the difference between computerized and paper creativity tests, this study employed a series of creativity tasks designed specifically for online use, but which could also be administered on paper so that the researcher could compare the results between the online and paper versions. The tests contained three types of tasks: *the Line Meaning test*, *the Drawing test*, and *the Real-world Problem test*.

The Line Meaning test was first used by Wallach and Kogan (1965) as a figural (or visual) test of creativity [17]. Usually, several incomplete and irregular figures were presented and test takers were asked to come up with as many meanings as they could for each figure. Three figures were used in the present study, including a curve and point figure, a wave-like figure, and an angle-like figure. First, an example (adapted from Wallach and Kogan's original test) was presented explaining how to respond to the test prompt. Then the participants were asked to imagine and write down all the things each figure might be.

The Drawing test had two parts: the *Story Drawing* test and the *Square Drawing* test. The first part – the Story Drawing test – was adapted from Urban's (2005) creativity test [18]. Urban has argued that traditional creativity tests such as the divergent thinking test focus too much on the quantitative aspects instead of the aspects of quality, so he proposed a drawing production test that asks participants to complete a drawing on the basis of some given figural fragments. This drawing production test was deeply rooted in the Gestalt tradition, which holds that a diagnostic for creativity should be holistic and gestalt-oriented – the whole is more than the sum of the parts. The Story Drawing test used in our study took the three figures presented in the previous task (the Line

Meaning test) and asked participants to combine these figures to create an interesting story. The second part of the Drawing test was the Square Drawing test, as adapted from Guilford's (1967) Sketches test [19]. Here, participants were shown three identical squares and then were asked to add in details to the squares to create original and interesting pictures.

The Real-world Problem test had two questions, both of which were selected from Runco's *Realistic Presented Problems*, a part of the *Runco Creativity Assessment Battery (rCAB).* (Note: The use of this test was permitted by Runco before the start of this study). This test was designed to address the concern that divergent thinking has a low correlation with creative achievement in the natural environment as opposed to the classroom [20]. In the Real-world Problem test used in the current study, participants were first given an example explaining how to respond to the questions. Then they were presented with two problematic situations regarding school and home, and were asked to come up with as many solutions as they could.

Two types of responses, verbal responses and drawing responses, were collected in the tasks. Objective rating of *fluency* and *originality* was used to score verbal responses. More specifically, the researcher used the top 20% scoring method [21], counting the number of the responses given by less than 20 percent of the sample to each question. The Drawing test, which included Story Drawing and Square Drawing, did not yield objective ratings of originality, because the researcher could not pool the drawings in order to find uncommon responses – each drawing was unique from an objective perspective. Therefore, subjective rating of *originality*, which is also called the average creativity method, was used to score drawing responses [22]. More specifically, five blind raters were recruited and trained to rate the drawing responses. The raters were instructed to use a 1-6 scale to rate each response for creativity, with "1" meaning "not creative at all" and "6" meaning very creative. To obtain the final score, the participant's ratings were summed and divided by the number of responses.

C. Online Assessment Interface

To build an online assessment system that allows participants to draw pictures, the researcher experimented with different web tools. The final solution involved several online services, some of which were free and others that charged a reasonable monthly fee.

1) Web drawing tool

The most important tool utilized in this study was a web drawing tool embedded in an online survey system. The tool, called A Web Whiteboard (AWW), is a web drawing interface designed for simple drawing and communication on computers and tablets. Two features of this tool are worth noting. First, AWW can be used as a plugin that can be embedded in most webpages. Therefore, the researcher was able to combine questions requiring verbal responses and those requiring drawing responses into one streamlined survey system. Second, AWW allows the customization of tools and other interface elements to be presented to participants. Thus, the study could utilize different testing materials for different tasks and create conditions for different online groups. With these features, AWW served as a powerful tool for the current study. Both the Story Drawing and Square Drawing tests used AWW as their drawing interfaces.

2) Online survey system

One critical feature that an online test should possess is the capability to streamline the testing procedure, helping test takers to complete all the questions without distraction or disruptions. An online survey service, as provided by Qualtrics through the link from the university, enabled the collection of responses through the internet. Several features were used to meet the needs of the current study. First, a simple link distributed via different email systems enabled the researcher to collect the required amount of data within a relatively short period of time. Second, a system randomizer helped the researcher assign participants to different groups based on the research design. Third, simple coding could be used to embed the drawing tool, as described above, into the survey, and the system also provided ways for participants to upload the pictures they had drawn. The only disadvantage of this service was the lack of an advanced coding environment, so that the drawing tool could not be *directly* embedded into the survey. Because of this, another service needed to be used, which will be described in the following section.

3) Web hosting service

AWW – the drawing tool – allows users to customize its interface to accommodate different needs; this was both an advantage and a disadvantage in the context of present study. While the advantage is obvious, the disadvantage is that it involves coding that is not recognized by the Qualtrics survey system. Thus, the researcher could not customize the drawing interface to create experimental conditions, including the presentation of test materials (instructions and figures) and control over what tools were shown to participants. To compensate for this issue, the researcher used a "bridge" that not only allowed the researcher to change the drawing interface but also allowed the drawing interface to be embedded into the streamlined Qualtrics survey system. Weebly, a web hosting service, played this role.

Two important features of Weebly helped the researcher connect the drawing tool with the survey system. First, Weebly allows and recognizes the HTML coding used to manipulate the web drawing interface. Second, Weebly can enable an SSL (Secure Sockets Layer) function that uses an encrypted link between the web server and browser, which in turn is recognized and accepted by the Qualtrics survey system.

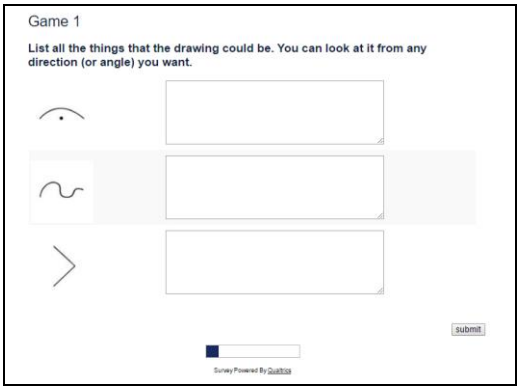

Fig. 1. Online test interface example 1 – The line meaning test.

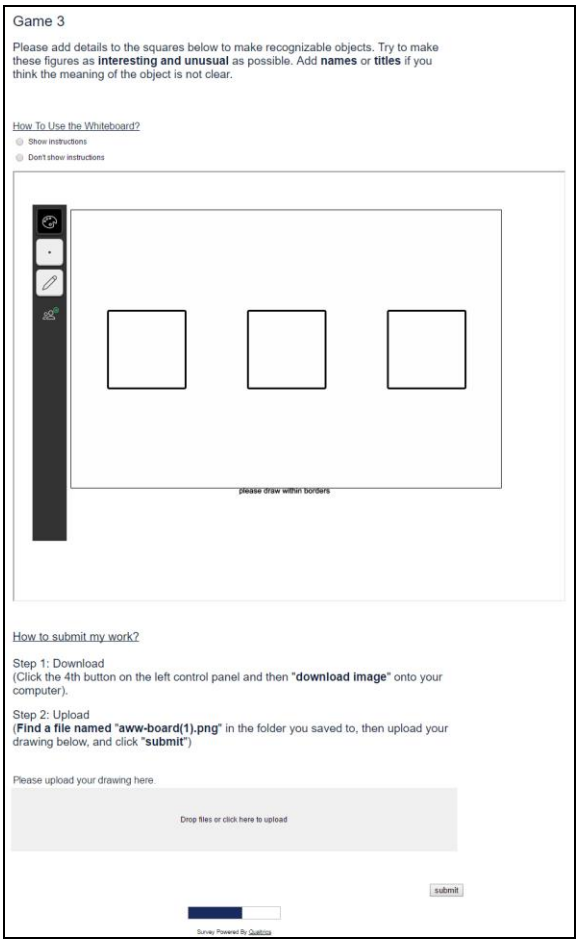

Fig. 2. Online test interface example 2 – The square drawing test.

After a few simple steps to connect AWW, Weebly, and Qualtrics, the entire online test system, which can collect both verbal and drawing responses, was successfully set up for the present study (**See Fig. 1 and Fig. 2 for examples**). (Note: Although participants can have access to the system via either PCs/Macs or tablets, results showed that only 1 out of 164 participants completed the test on a tablet.)

D. Procedures

After interested participants received an email or saw an advertisement posted on a school announcement board, the test/survey link brought them to a page giving information about the present study. After reading the information sheet and clicking the "agree" button, they were randomly assigned into three groups by the system randomizer and received different instructions accordingly.

1) Paper and pencil (PP) group

For the paper and pencil (PP) group, participants were directed to an online ticket platform (Eventbrite), where they were asked to select the times they could attend lab sessions to complete the paper version of the creativity test and the two surveys. The sessions were held in a lab located in a university building. The elements of informed consent were explained when participants arrived at the lab for their sessions and the participants were provided an informed consent form. After they had signed the form, they received the instructions for the tasks. To maximize each participant's performance and reduce potential anxiety associated with test taking, the investigator attempted to create a "game-like" environment by saying that this is a thinking game. The general instructions were as follows:

"This is a thinking game. All the instructions are on the sheets. Please follow these instructions and complete all the tasks and questions, taking your time. If you have any questions, let me know."

The participants were then provided with a pencil with an eraser and started the test. After finishing the tasks, each participant received 10 dollars as compensation, and the procedure was over for the group. All paper-and-pencil responses were transcribed into electronic versions.

2) Online groups

For the online basic (OB) group and the online advanced (OA) group, all the tasks, including the creativity test and the two technology surveys, were administered online. The participants followed the instructions posted on the (Qualtrics) survey that was designed specifically for the current study.

The difference between the online basic (OB) group and the online advanced (OA) group was that, in the figural tasks of the creativity test, the OB group was asked to use only two drawing tools in the editor (pencil and eraser) to complete the figural tasks, while OA group was allowed to use more colors and brush sizes for the tasks. Upon finishing the final task, the procedure ended for both online groups. The participants were then automatically directed to the online ticket platform (Eventbrite), where they could select times to pick up their 5 dollar compensation or choose to receive their compensation in the form of an Amazon e-gift card.

3) Raters

Five raters were recruited from the university to rate the creativity of each response. The investigator held an approximately 30-minute video meeting with each rater to explain the scoring procedure. After the meeting, the raters received an email re-iterating the general instructions explained in the meeting, along with several Excel files (containing the verbal responses) and PowerPoint files (containing the drawing responses), with specific instructions for each subtest. After scoring was completed, the raters sent the files back. Each rater received $100 for completing the scoring.

III. Results

A. Reliability Analysis

An important issue to consider before further analysis is the comparability of the online and paper-and-pencil versions of the creativity test in terms of reliability estimates. If large differences existed between different groups in terms of reliability coefficients, there would be no need to compare scores between groups. Therefore, a reliability analysis of the six creativity scores was conducted for each group, as summarized in Table I. Specifically, to assess internal consistency of the scores obtained from the Line Meaning test and the Real-world Problem test, Cronbach's alpha and the Spearman-Brown (S-B) coefficient were used. Cronbach's alpha is often calculated for a test that has three or more items, while the S-B coefficient is most appropriate when a test has only two items [23].

For the Drawing test, the inter-rater reliability of each item was assessed by interclass correlation coefficients (ICCs).

TABLE I: SUMMARY OF RELIABILITY INFORMATION OF THE SIX CREATIVITY SCORES FOR DIFFERENT GROUPS

	Online-Basic Group	*Online-Advanced Group*	*Paper-and-Pencil Group*	*Overall*
Index		*alpha*		
Line Meaning fluency	0.91	0.93	0.94	0.93
Line Meaning originality	0.81	0.90	0.89	0.88
		ICC		
Story Drawing originality	0.69	0.86	0.85	0.81
Square Drawing originality	0.90	0.91	0.90	0.91
		Spearman-Brown coefficient		
Real-world Problem fluency	0.78	0.79	0.86	0.81
Real-world Problem originality	0.69	0.78	0.63	0.73

According to the results, all the reliability coefficients were within the acceptable range. Furthermore, the differences between groups in terms of reliability coefficients were acceptable as well, indicating that scores obtained from the online groups were generally as reliable as the scores obtained from the paper-and-pencil group.

B. Differences in Creativity Scores between Paper-and-Pencil and Online Versions

1) Group and gender effects on each creativity score

A series of two-way factorial analyses of variance (ANOVAs) were performed to establish and explore group and gender differences with respect to each creativity score. It was hypothesized that the advanced online group would achieve higher originality scores in the Drawing test than the basic online and paper-and-pencil groups. It was also hypothesized that gender may influence scores obtained on verbal tasks, including the Line Meaning test and the Real-world Problem test. The interaction between group and gender were also explored in this analysis.

Results from two-way ANOVA showed that males scored significantly higher than females on Line Meaning originality (F (1, 158) = 4.77, p = 0.030, η^2 = 0.03) and Real-world Problem originality (F (1, 158) = 4.11, p = 0.044, η^2 = 0.03). No other significant effects, including group effects (F (2, 158) ranges from 0.05 to 1.80, p ranges from 0.15 to 0.57) and interaction effects, were found for the creativity scores. These results indicate that the testing environment, whether online with different tools or in a classroom with pencil and paper, did not significantly affect the creativity scores.

2) Overall group effect on creativity scores

To test the overall effect of the three groups on creativity scores, MANOVA was conducted with the six creativity scores as dependent variables, and group as the independent variable.

Homogeneity of variance was examined at both the multivariate and univariate levels. Neither Box's M (p = 0.086) nor Levene's tests (p = 0.054 to 0.515) were significant. The multivariate effect of group difference on creativity scores was also not statistically significant, with Wilk's lambda = .88, F (12, 312) = 1.67, p = 0.073, η^2 = 0.06.

3) Overall group effect on creativity scores after controlling for the gender effect.

As creativity scores may vary by gender, to test the overall effect of three groups on creativity scores and control for the gender effect, MANCOVA was performed with the six creativity scores as dependent variables, group as the independent variable, and gender as the covariate.

Homogeneity of variance was examined at both the multivariate and univariate levels. Neither Box's M (p = .086) nor Levene's tests (p = 0.051 to 0.404) were significant. The multivariate effect of group difference on creativity scores was also not statistically significant, with Wilk's lambda = .90, F (12, 310) = 1.47, p = 0.133, η^2 = 0.05. Gender (the covariate) had no statistically significant effect on creativity scores either, with Wilk's lambda = .93, F (6, 155) = 2.05, p = 0.062, η^2 = 0.07.

IV. DISCUSSION

To evaluate the feasibility of an online creativity test, two research questions were proposed, including question 1 (*How reliable are the scores obtained from the creativity tests?*) and question 2 (*How do the results obtained from an online creativity test differ from the results of the paper-and-pencil version?*). Research Question 1 addresses the reliability aspect of the test, while Research Question 2 represents an effort to provide some evidence of validity for the test. The results for research questions 1 and 2 will be discussed to help us understand the affordances provided by technology in creativity education.

Two themes emerged while comparing creativity scores between groups. Specifically, the comparison of creativity scores showed no differences between the basic-online group, the advanced-online group, and the paper-and-pencil group after controlling for gender. This study also compared and found significant differences in creativity scores between male and female participants.

A. Summary of the Findings

1) Theme 1: No mode effect was found for creativity scores or their reliability estimates

a. No mode effect on reliability estimates

The current study found no dramatic differences in either inter-item or inter-rater reliability estimates between the three groups (online-basic, online-advanced, and paper-and-pencil). Although few studies have compared the reliability information of drawing tasks, our findings were consistent with previous research that compared reliability information for verbal responses between computer and paper versions of creativity tests. For example, Lau and Cheung (2010) compared the electronic and paper-and-pencil versions of the Wallach-Kogan Creativity Tests in terms of their internal consistency (alpha), and found that the magnitudes were

comparable. The tests they used only produced verbal responses. The present study has demonstrated that drawing responses can produce similar or even the same inter-rater reliability estimates across modes.

b. No mode effect on creativity scores

Contrary to the hypothesis, we found no differences in any of the creativity scores after the gender effect was controlled for. The results are consistent with the latest research on computerized creativity tests (Zabramski, Gkouskos, & Lind, 2011), while contradicting findings obtained 20 years ago by Kwon (1994). Initially, our hypothesis was made on the basis of the idea that mouse input would have a negative impact on participants' drawing performance. As Zabramski and Neelakannan (2011) have claimed, feedback from mouse movements is presented on the computer screen, which is spatially separated from the user, making the entire human-computer interaction indirect. In addition, mouse movement is usually smaller than cursor movement on the screen, which would also affect drawing accuracy in the test. However, neither the present study nor the study of Zabramski and colleagues found significant differences in creativity scores between different input methods, indicating that a mode effect is not present in creativity testing.

An important factor that might contribute to the inconsistency between the results from now and two decades ago is computer familiarity. People who use computers often may develop motor skills that would compensate for the inaccuracy and inconvenience that comes with the use of a mouse. What is evident about the changes in computer use across time is that people nowadays use computers and mice more often than people living 20 years ago. According to the 2014 U.S. Census, more than 80% of American households now own a computer, compared to 30% twenty years ago. College students, a major component of our sample, would have even more access to computers. Zabramski and colleagues collected their data in Sweden, a country that has the most computers per household in the world [24]. Their sample consisted of adults aged from 20 to 38. In contrast, the sample in Kwon's (1996) study consisted of fifth and sixth graders. Therefore, both time and age would lead to differences in participants' computer familiarity, which would in turn affect their performance in computerized drawing tests.

Interestingly, Zabramski and colleagues (2013) provided a different explanation, arguing that the graphic user interface (GUI) might lead to additional cognitive load, influencing participants' drawing performance [25]. They believed that the more elaborate GUI used in Kwon's (1996) study – as opposed to the simplified GUI used in their study – was the major factor contributing to the differences between the computerized and paper versions. This claim, however, is not supported by the findings of the present study. The present study used elaborative GUIs in both the basic-online group and advanced-online group, according to Zabramski's standards (drawing tools were visible on the screen), but no differences were found between these two online groups and the paper-and-pencil group. Neither did Zabramski and his colleagues (2013) find any differences in creativity scores between the elaborate and simple GUIs.

2) Theme 2: Gender differences in creativity scores were inconsistent

Significant gender differences were found in the Line Meaning originality scores and the Real-world Problem originality scores, with males producing significantly higher overall mean scores than females. The results added more mystery to the mixed evidence on gender differences found in the previous literature. In fact, no consistent pattern of gender differences in creativity tests has been identified [26], [27], with a greater number of studies finding that girls or women score higher on verbal creativity tests.

One possible explanation for the gender effect in the present study is that the participants were self-selected. Self-selected males might produce higher originality scores than females in verbal tasks. However, the results of the present study should be interpreted with caution, as there might be bias due to the fact that the females outnumbered the males by almost double. Therefore, further research needs to be done to study gender differences in creativity.

B. Summary of the Feasibility of an Online Creativity Assessment System

The above analyses have yielded important implications for the use of an online creativity assessment system. All of the issues discussed above, together with the design issues of the online testing system, will be summarized and further discussed below.

1) Limitations and disadvantages

Several limitations should be noted before we discuss the potential advantages of an online creativity test. The present study represents an effort to computerize the creativity test using current available technologies. One of the goals was to develop and test a system, rather than to publish a new tool. To achieve this goal, compromises needed to be made on several aspects.

First, the current study represents one of the first steps (for example, ANOVA) to test for the differences between the test versions. To further establish the equivalence between these versions, a much bigger sample and more sophisticated analyses (such as DIF analysis within IRT modelling and measure variance) are needed in the future. In another word, despite that the present study found no differences, it only compares the mean performance. More research is needed to investigate the differences in more aspects.

Second, unlike the application development done in the previous studies, none of the technologies involved in the present study were owned by the investigator. For example, the web drawing interface was designed and maintained by a third-party. The investigator only added some coding to adapt the tool into the testing system. One of the downsides of this method was that the researcher was unable to have total control of the tool if any errors occurred. In fact, in the initial experimental design, the drawing interface presented to the advanced-group contained more tools, such as a straight-line burton and circle burton, to allow participants more control during the drawing process. In the actual experiment, however, these tools mysteriously disappeared, despite communication between the investigator and the tool's developer. Neither could identify the cause the problem, leading to fewer tools being shown in the drawing interface for the advanced group.

Third, an online creativity test may not be suitable for small-scale assessments. The findings in the present study have demonstrated that the online and paper versions were equivalent in almost every respect, which leads to the next question: If they are equivalent, why would we use an online test? In other words, is there value added by putting the test online or computerizing it? The answer may be no. For example, a small-scale and occasional assessment, which can usually be carried out in a classroom, does not need to be computerized. Using an online version would not only be an expensive option, but would also bring with it technological issues.

2) Advantages and potential of online creativity assessment system

With the aforementioned limitations in mind, the present study also demonstrated several advantages and the potential of an online creativity test.

First, online testing can reach more people within a relatively short amount of time, at a lower cost. Paper-and-pencil testing usually requires people to be at a particular place at a certain time to take the test. If people's schedules conflict with the test time, they will need to change their schedule, or the test administrator will need to change the time or place. This issue does not apply to online testing. As long as the test takers have the link to the testing system, they can choose to do the test at the time and place that suits them. This is especially an important feature when test administrators want to collect self-selected samples in a large-scale assessment.

Second, current technologies allow test designers to develop their own creativity testing systems. Despite the pitfalls mentioned in the previous section, there are several advantages of using third-party technologies. For example, test designers do not need to possess advanced technological knowledge (such as programming or web design skill) to develop such a tool, thus reducing learning time as well as any costs associated with hiring technology assistants. In addition, since different interfaces do not significantly influence creativity test performance (including both drawing and verbal performance), test designers/administrators can experiment with different tools to determine which solutions best satisfy their purposes.

Third, technologies hold great potential in allowing people to develop automatic and adaptive creativity tests. In the present study, objective scoring of verbal responses produced high reliability estimates. This has important implications for automatic scoring because objective scoring, which is carried out by first pooling all the responses and then counting the number of responses given by less than 20% of the sample, can be conducted via computer program if it can recognize the meaning of the responses. Given that more and more artificial intelligence (AI) systems such as SIRI, Alexa, and Cortana, can nowadays easily recognize and respond to simple commands from people, it would not be very difficult to design a computer program that recognizes verbal responses to various types of creativity tasks. If automatic scoring is feasible, then adaptive creativity testing is also possible. This of course requires more research in the future.

In summary, the findings of this study indicate that an online creativity assessment system can produce the same, reliable creativity scores as the paper test does. This form may be more useful for large-scale assessments because it makes data collection much faster and more convenient. The current study also demonstrates that online creativity testing could add more value if automatic scoring and adaptive testing are used.

REFERENCES

[1] T. M. Amabile and J. Pillemer, "Perspectives on the social psychology of creativity," *The Journal of Creative Behavior*, vol. 46, no. 1, pp. 3–15, Mar. 2012.

[2] R. Florida, "High-tech innovation creativity and regional development," in *Creativity and Innovation: Theory, Research, and Practice*, J. A. Plucker, Ed. Waco, TX: Prufrock, vol. 1, 2017, pp. 61-74.

[3] J. M. George, "Creativity in organizations," in *The Academy of Management Annals*, J. P. Walsh and A. P. Brief, Eds. vol. 1, New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates, 2008, pp. 439-477.

[4] M. D. Mumford and B. Licuanan, "Leading for innovation: Conclusions, issues, and directions," *The Leadership Quarterly*, vol. 15, no. 1, pp. 163–171, Feb. 2004.

[5] P. Tierney, S. M. Farmer, and G. B. Graen, "An examination of leadership and employee creativity: The relevance of traits and relationships," *Personnel Psychology*, vol. 52, no. 3, pp. 591–620, Sep. 1999.

[6] M. C. Kwon, "An exploratory study of a computerized creativity test: Comparing paper-pencil and computer-based versions of the Torrance tests of creative thinking," Ph.D. dissertation, Texas A&M Univ, College Station, TX, 1996.

[7] J. E. Pretz and J. A. Link, "The creative task creator: A tool for the generation of customized, web-based creativity tasks," *Behavior Research Methods*, vol. 40, no. 4, pp. 1129–1133, Nov. 2008.

[8] S. Lau and P. C. Cheung, "Creativity assessment: Comparability of the electronic and paper-and-pencil versions of the Wallach–Kogan Creativity Tests," *Thinking Skills and Creativity*, vol. 5, no. 3, pp. 101–107, Dec. 2010.

[9] A. K. Palaniappan, "Web-based creativity assessment system," *International Journal of Information and Education Technology*, pp. 255–258, 2012.

[10] S. Zabramski, "Creating digital traces of ideas. Evaluation of computer input methods in creative and non-creative drawing. Storytelling," Ph.D. dissertation, Uppsala: Acta Universitatis Upsaliensis, 2014.

[11] P ásztor, G. Molnár, and B. Csapó, "Technology-based assessment of creativity in educational context: The case of divergent thinking and its relation to mathematical achievement," *Thinking Skills and Creativity*, vol. 18, pp. 32–42, Dec. 2015.

[12] R. W. Hass, "Feasibility of online divergent thinking assessment," *Computers in Human Behavior*, vol. 46, pp. 85–93, May 2015.

[13] H. V. Leeson, "The mode effect: A literature review of human and technological issues in computerized testing," *International Journal of Testing*, vol. 6, no. 1, pp. 1–24, Mar. 2006.

[14] S. Zabramski, D. Gkouskos, and M. Lind, "A comparative evaluation of mouse, pen-and touch-input in computerized version of the Torrance tests of creative thinking," in *Proc. the 1st European Workshop on HCI Design and Evaluation*, pp. 57-61, 2011.

[15] S. Zabramski and S. Neelakannan, "Paper equals screen: a comparison of a pen-based figural creativity test in computerized and paper form," in *Proc. the Second Conference on Creativity and Innovation in Design*, pp. 47-50, 2011.

[16] L. A. Jackson, E. A. Witt, A. I. Games, H. E. Fitzgerald, A. Eye, and Y. Zhao, "Information technology use and creativity: Findings from the children and technology project," *Computers in Human Behavior*, vol. 28, no. 2, pp. 370–376, Mar. 2012.

[17] M. Wallach and N. Kogan, *Modes of Thinking in Young Children*, New York: Holt, Rinehart, & Winston, 1965.

[18] K. K. Urban, "Assessing creativity: The test for creative thinking-drawing production (TCT-DP)," *International Education Journal*, vol. 6, pp. 272-280, 2005

[19] J. P. Guilford, *The Nature of Human Intelligence*, New York, NY: McGraw-Hill, 1967.

[20] S. M. Okuda, M. A. Runco, and D. E. Berger, "Creativity and the finding and solving of real-world problems," *Journal of Psychoeducational Assessment*, vol. 9, no. 1, pp. 45–53, Mar. 1991.

[21] J. A. Plucker, M. Qian, and S. L. Schmalensee, "Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking," *Creativity Research Journal*, vol. 26, no. 2, pp. 135–143, Apr. 2014.
[22] P. J. Silvia, B. P. Winterstein, J. T. Willse, C. M. Barona, J. T. Cram, K. I. Hess, J. L. Martinez, and C. A. Richard, "Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods.," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 2, no. 2, pp. 68–85, May 2008.
[23] R. Eisinga, M. Grotenhuis, and B. Pelzer, "The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?" *International Journal of Public Health*, vol. 58, no. 4, pp. 637–642, Oct. 2012.
[24] OECD, *OECD Factbook 2015-2016: Economic, Environmental and Social Statistics*, Paris, France: OECD Publishing, 2016.
[25] S. Zabramski, V. Ivanova, G. Yang, N. Gadima, and R. Leepraphantkul, "The effects of GUI on users' creative performance in computerized drawing," in *Proc. the International Conference on Multimedia, Interaction, Design and Innovation*, pp. 142-151, 2013.
[26] J. Baer and J. C. Kaufman, "Gender differences in creativity," *The Journal of Creative Behavior*, vol. 42, no. 2, pp. 75–105, Jun. 2008.
[27] P. C. Cheung and S. Lau, "Gender differences in the creativity of Hong Kong school children: Comparison by Using the new electronic Wallach–Kogan creativity tests," *Creativity Research Journal*, vol. 22, no. 2, pp. 194–199, May 2010.

J. Guo was born in Shanghai, China. He went to study in the US in 2012 and earned his doctoral degree in educational psychology from the University of Connecticut in 2016.

He is currently working as a postdoctoral research fellow at the East China Normal University, Shanghai, China. He was a senior research fellow at the Johns Hopkins University Center for Talented Youth. He has published articles on the *Roeper Review*, *Creativity. Theories – Research – Applications*, and *Thinking Skills and Creativity*. His areas of interest include developing creativity measurement instruments and assessment tools, talented and gifted education, development of creative potentials, teaching creativity in the classroom, and use of technology in creativity enhancement.

Dr. Guo is a member of American Psychological Association (APA) Division 10 - Society for the Psychology of Aesthetics, Creativity and the Arts.