An intelligent scoring method for a creative thinking test

Yeni Anistyasari†, Ekohariadi†, IGP Asto Buditjahjanto† & Shintami C. Hidayati‡

State University of Surabaya, Surabaya, Indonesia†
Sepuluh Nopember Institute of Technology, Surabaya, Indonesia‡

ABSTRACT: A well-known test for assessing creativity is the Torrance Test of Creative Thinking Figural (TTCT-F). It includes scores for fluency, flexibility, originality, elaboration and abstractness of titles in the figural version. These assessment measures are highly subjective and largely dependent on the assessors' analysis and knowledge. Also, the evaluation requires much effort in efficiency and cost. Thus, this article presents an efficient way for evaluating creative thinking that is both consistent and meaningful, particularly on the originality scale which refers to the rarity of ideas. The originality scales are calculated using a sparse coding-based scale-invariant feature transform (ScSIFT) algorithm. The proposed method is evaluated using the TTCT-F administered to 202 students. Sparse dictionary learning, sparse query image processing and picture matching algorithms were employed to process student responses. The effectiveness of this proposed method was evaluated by comparing it to a manual assessment of TTCT-F by expert judgments. The results revealed that the proposed method is as accurate as expert judgments. However, the proposed method saves more time and is more objective.

INTRODUCTION

Following the rising disruption caused by technology in the 21th Century workplace, numerous experts and practitioners highlighted creativity as the most crucial trait for successful workplace [1][2]. While scholars have noticed the impact of technology on creative teaching practices in classrooms over the last decade, just a few studies have focused on the use of technology to measure creativity. Moreover, most creative evaluations are still administered using paper and pencil examinations, rendering them inappropriate for new objectives, such as large-scale testing and speedier data collecting [1-3].

One of the earliest and most important parts of creativity research is the study of divergent thinking [3][4]. Divergent thinking is the most promising basis for creative ideas, according to a psychometric examination of creativity [5].

Torrance Tests of Creative Thinking (TTCT) are often used to test divergent thinking because of this. To begin with, this test was perfect: it was simple to administer and basic to score. The test was built by Torrance based on Guilford's work on divergent thinking [6][7].

These instruments of evaluation include linguistic and figurative elements. Verbal-based items include both stimuli and reactions that are expressed verbally. Figural-based objects have stimuli that are figural, yet the reaction can be verbal or figurative. TTCT items can be graded on a variety of scales, including fluency, flexibility and originality.

Fluency is a term that describes one's ability to come up with numerous ideas in response to a certain problem. It is measured by how many thoughtful, suitable comments are made by the tested. To be flexible, one must be able to look at a situation from several perspectives and be able to categorise reactions into different groups.

A person's originality is defined as their ability to come up with new, original ideas, and it is usually measured by the statistical infrequency with which those ideas appear within a given sample. In spite of this, emerging research on creativity shows that divergent thinking activities have a wide range of applications. It is important to perform more repetitions if the sample size is higher [8].

In education, technology-based assessment research is one of the fastest growing segments [9-11]. There is an ever-increasing interest due to the advantages of technology-based assessments, such as the administration of tests on-line and the automated scoring and exact feedback. Existing assessment instruments are difficult, if not impossible, to use in educational settings because of their complexity or the difficulty of scoring [12].

Additionally, creativity tests, such as the TTCT, require the responses to be evaluated by a trained individual, but this preference is not always followed. Such an assessment is time-consuming, costly and subjective. Creativity scores can vary significantly between reviewers and are highly personal and dependent on the reviewer's interpretations and knowledge.

These findings are highly inconclusive. As a result, the authors of this article propose a novel method for utilising technology to develop an automated assessment tool for creative thinking that will partially score the TTCT Figural of (TTCT-F) originality scale. The originality scale for creative thinking abilities is identical to matching multiple images and calculating their similarity. The lower the score for similarity, the higher the score for originality. Scale-invariant feature transform based on sparse coding (ScSIFT) is used in this study to compute similarity [13].

METHODS

Data Acquisition

The Torrance Tests of Creative Thinking are used in this study, based on the Figural form (TTCT-F). Two-hundred and two students were involved in creating digital storytelling using Scratch 3.0 visual programming. The students' responses were then analysed to calculate their originality score in terms of creative thinking. The evaluation was conducted using two techniques: the proposed intelligent scoring method and expert judgments in creative thinking assessment. The two techniques were then compared using the student's *t*-test to see a statistically significant difference.

Intelligent Scoring Methods

The ScSIFT algorithm is the extension of the SIFT feature extracted from keyframe segments as a training dataset for the over-complete dictionary to assemble a collection of over-complete sources [11]. The query image's SIFT vector is sparsely encoded using the over-complete dictionary. After that, the sparse vector is indexed. Furthermore, the key-SIFT image's feature index is compared to the query image's feature index using the over-complete dictionary to generate a set of comparable candidate sets, matching the key-image sparse coefficient to the query image sparse coefficient to capture similar image detection results.

The image the ScSIFT features are further defined as the SIFT feature sparse coding of the sparse coefficient vector. The image similarity method used to calculate the creative thinking original score consists of three stages: sparse dictionary learning, sparse query images and image matching algorithms. Each algorithm's pseudocode is presented in Figure 1 to Figure 3. The interface of intelligent scoring system is depicted in Figure 4.

Algorithm 1 - Sparse dictionary learning

```
n ← number of images from query library
for (k = 1; k <= n; k++)
  Training_Fk ←extract SIFT features of image Ik
endfor
normalize Fk according to Eq. (1)
train the overcomplete dictionary D using Training_Fk
output dictionary D
```

Figure 1: Sparse dictionary learning.

Algorithm 2 - Query image sparse

```
s = 1
while I is uncoded image in the query library
do
  read image Is of query library
  image_Feature_Fs ←extract SIFT features of image Is
  normalize Fs according to Eq. (1)
  ScSIFT ← preserve the sparse coefficient α
  s++
endwhile
create index
save ScSIFT in the query library
```

Figure 2: Query image sparse.

Algorithm 3 - Image matching

```
while column ≠ maximum
do
   Read the image I
   Fs ← extract its SIFT feature
   Normalize Fs
   ScSIFT feature β ← Get the sparse representation of
   the normalized feature set Fs by the sparse algorithm
   σ ← ScSIFT feature similarity distance threshold
   θ ← image similarity threshold
   read the first column of ScSIFT feature x = β0.
   Search for the nearest k column coefficients of the
   ScSIFT feature x
   Calculate the distance dk of α according to Eq.(6)
endwhile
calculate the total number of sparse feature points
calculate the similarity degree
if similarity degree ≠ 0
    Originality score = 0
endif
```

Figure 3: Image matching.

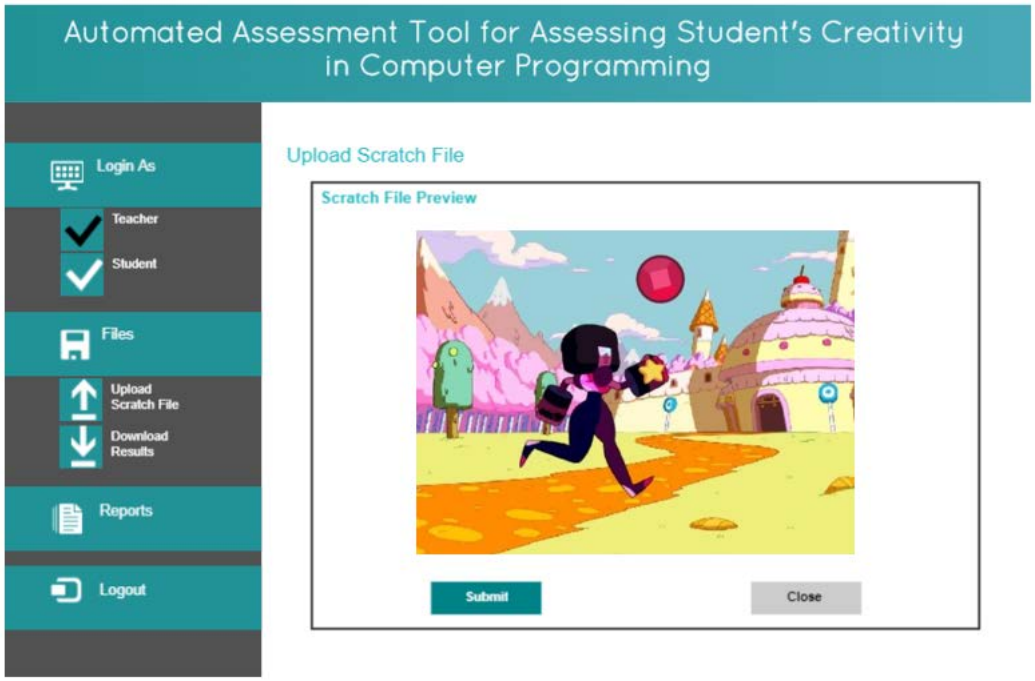

Figure 4: Interface of the intelligent scoring system.

RESULTS AND DISCUSSION

Results of Intelligent Scoring Method for Creative Thinking Test

Students were requested to create digital storytelling. Their responses were then evaluated using the proposed intelligent scoring method. The results were converted to scale 0-5 (Table 1).

Table 1: The conversion from ScSIFT results to originality scale of creative thinking.

ScSIFT results	Originality score
0.76 - 1.0	0
0.61 - 0.75	1
0.46 - 0.6	2
0.31 - 0.45	3
0.16 - 0.3	4
0 - 0.15	5

The student's response receives a score of zero if it is comparable to other students' responses, but one if it is unique and distinct from other students' responses. As a result, each student's maximum score is 5. The frequency of each score is depicted in Figure 5. The number of students who obtained score 0, 1, 2, 3 and 4 were 36, 40, 51, 40 and 35, respectively. There was no student with score 5 (Figure 5).

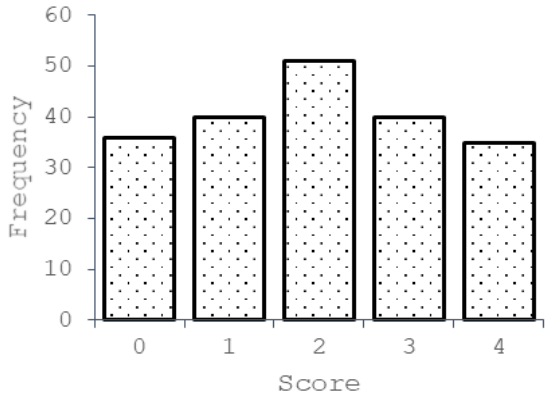

Figure 5: Creative thinking score according to the proposed method.

Results of Expert Judgments for the Creative Thinking Test

Students' responses of digital storytelling was simultaneously evaluated by the expert in creative thinking. Students were awarded 0 if their responses were similar to others. Score 5, 4, 3, 2 and 1 was respectively obtained if 2, 4, 6, 8 and 10 characters were the same as others. The frequency of each score is depicted in Figure 6 below. The number of students who obtained score 0, 1, 2, 3 and 4 were 35, 41, 48, 44 and 34, respectively. Similarly to the proposed method results, there was no student with score 5 (Figure 6).

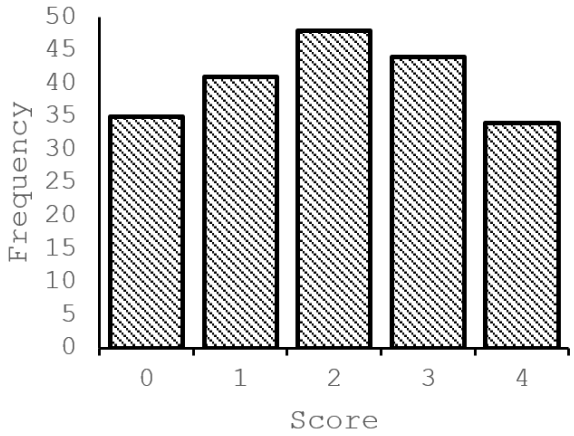

Figure 6: Creative thinking score according to the expert judgments.

Comparison of the Intelligent Scoring Method and Expert Judgments Results

The number of students for each score results is listed in Table 2. In the study, accuracy of the proposed method was examined. Accuracy refers to the degree to which a value is near to its true value. First, a confusion matrix was built as shown in Table 3. Accuracy is the proportion of true results which is formulated as *true positive score* divided by *total score*. Hence, the accuracy of intelligent scoring is 87%. This score indicates that the accuracy is high.

Table 2: Number of students for each score.

Score	Number of students	
	Intelligent scoring	Expert judgments
0	36	35
1	40	41
2	51	48
3	40	44
4	35	34
5	0	0

Table 3: Confusion matrix.

		Expert judgments	
		0-2	3-5
Intelligent scoring	0-2	111	12
	3-5	14	65

A statistical analysis, two-tailed student *t*-test, was then exploited to evaluate the significance different between the results of the intelligent scoring method and expert judgments. The hypothesis is formally written as follows:

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Where μ_1 is the average of intelligent scoring results and μ_2 is the average of expert judgments results. The level of significance, α, is 0.05. The two-tailed *t* and *p* values equal to 0.2794 and 0.7622, respectively. Since $0.7662 = p \geq \alpha = 0.05$, H_0 is kept. This indicates, the average of intelligent scoring results is the same as the average of expert judgments results. The results of two tailed *t*-test are listed in Table 4.

Table 4: Two-tailed student *t*-test results.

	Intelligent scoring	Expert
Sizes	202	202
Means	1.995	1.9554
Standard deviation	1.394	1.4534
Degrees of freedom	401.3187	
Critical *t* value	1.96589	
95% confidence interval	-0.239 ; 0.3182	
t statistic	0.2794	
p value	0.7622	

Furthermore, the expert who analysed the TTCT-F stated that manually assessing the responses of 202 students was a waste of time. Each response took about 15 minutes to get examined or compared to other responses. In total, it required 3,030 minutes or 50.5 hours. In comparison, the proposed method took about 0.5 second to evaluate all responses. The total time for examining 202 was 101 seconds.

Discussion

For a long period of time, creativity evaluation depended on subjective judgments of assessors to determine the originality of ideas and products. While manual scoring systems have been beneficial in the field, they have two significant drawbacks - labour cost and subjectivity - that expose dependability and function as a barrier for researchers without the capacity to code thousands of replies. In this study, the researchers attempted to solve these shortcomings of subjective scoring by using recent advances in automated creativity evaluation via artificial intelligence. They established that the proposed intelligent system accurately predicts expert judgments on a commonly used TTCT-F. Additionally, the accuracy of intelligent scoring in comparison to expert judgments in analysing the TTCT-F using the consensus evaluation approach is essential. A previous study was based on manually assessing the TTCT-F scores before submitting them to six assessors (three psychologists and three artists). The most that these judges could do with those scores was to obtain inter-rater reliability values less than 0.78; some studies ignore this accuracy measurement.

The most essential is the effort savings enabled by the proposed method. It takes approximately 15 minutes for the expert to grade a single TTCT-F response. It is instantly obvious that even modest investigations utilising the TTCT-F need considerable work. In comparison, 1,887 participants would require 471,629 hours of labour for a single rater. In the current study, the scoring process required 50.5 hours equal to seven work days. Assuming the labour wages of 25 USD/day, as is the case in Indonesia, 175 USD would be required to examine 202 TTCT-F responses of digital storytelling.

It is critical to understand the constraints of ScSIFT while assessing creativity. While the ScSIFT is a valuable tool for studying the originality scale of creativity, it is not always more trustworthy or more valid than subjective judgments. In this vein, the authors of this article discovered that human judgments had a numerically greater degree of validity when compared to correlations with other creative measures. Notably, they discovered that the ScSIFT is a valid substitute for human judgments. Another distinguishing property of ScSIFT is its close connection to human assessments of novelty versus creativity. Indeed, the findings indicate that the ScSIFT is slightly more sensitive to novelty than to creativity, which is consistent with the similarity-based techniques that were employed to calculate these values. Because both humans and the intelligent scoring method are sensitive to conceptual remoteness, a similarity distant response is likely to be seen as unique by humans. However, the creative criterion carries the additional weight of utility, i.e. whether the response is appropriate, humorous or brilliant, which similarity distance cannot yet convey.

Finally, similarity distance is a novelty metric, not a direct path to creativity, but a demonstrably valid substitution. At the same time, the authors would suggest that undergraduate students, who often assess responses to creativity activities, are not a reliable predictor of creativity. Indeed, previous research has revealed flaws with their statistics as well (e.g. fatigue, bias, disagreement, etc). Furthermore, because studies on creativity disagree on what constitutes a creative concept, both similarity distance and human raters may be flawed, although in different ways. Finally, considering the costs associated with subjective human evaluations, it appears that even if automated assessments get near to the levels of validity associated with human ratings, this is a significant step forward.

Limitations

A significant shortcoming of the intelligent scoring method approach employed in this study is that the TTCT-F scores transition from continuous (0-1), which is generated from the ScSIFT algorithm to categorical (0-5) values. There are two practical consequences of this change in the existing method. The responses might be classified as the same as expert judgments and fall at the cost of gaps between the scoring conversions range. It is also possible to lose the ability to provide highly personalised diagnostic feedback. With category automatic scoring, feedback is confined to more generic suggestions for each range. While the automatic scoring does not exclude the provision of comprehensive feedback, the work required negates some of the automated scoring's benefits.

CONCLUSIONS

Quantifying a student's creative thinking abilities, such as originality, is time-consuming. This study offers an automated evaluation method for the Torrance Tests of Creativity Figural (TTCT-F). The evaluation demonstrates that the suggested approach outperforms manual evaluation. The *t*-test findings show that there is no statistically significant difference. Furthermore, the proposed approach is highly accurate and faster than the manual method. As is the case with the majority of AI applications across a range of fields, the true goal of this technology is not to replace humans, but to liberate them from monotonous and unproductive tasks. Other dimensions of creative thinking, such as fluency, flexibility and elaboration, are proposed for further investigation in the automated assessment instrument.

ACKNOWLEDGEMENTS

The current study was supported by the State University of Surabaya (Universitas Negeri Surabaya) Research Funding 2021.

REFERENCES

1. Harsiati, T., Pradana, I.M.P. and Amrullah, H., Information literacy and self-regulation in the context of the creative thinking of prospective engineers. *World Trans. on Engng. and Technol. Educ.*, 17. **2**, 197-203 (2019).
2. Malik, A. and Ubaidillah, M., Students critical-creative thinking skill: a multivariate analysis of experiments and gender. *Inter. J. of Cognitive. Research in Science, Engng. and Educ.*, **8**, 49-58 (2020).
3. Avsec, S. and Ferk Savec, V., Creativity and critical thinking in engineering design: the role of interdisciplinary augmentation. *Global J. of Engng. Educ.*, 21, **1**, 30-36 (2019).
4. Guilford, J.P., *Fundamental Statistics in Psychology and Education*. Wiley, 41, **3** (1957).
5. Silvia, P.J., Winterstein, B.P., Willse, J.T., Barona, C.M., Cram, J.T., Hess, K.I., Martinez, J.L. and Richard, C.A., Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2, 68-85 (2008).
6. Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E. and Van den Noortgate, W., The factor structure of the figural torrance tests of creative thinking: a meta-confirmatory factor analysis. *Creative Research J.*, 30, **4**, 352-360 (2018).
7. Ng, A.W.Y. and Lee, C.-Y., Assessment of creative thinking of Hong Kong undergraduate students using the torrance tests of creative thinking. *Proc. Inter. Conf. on Higher Educ. Adv.*, Valencia, Spain, 1-8 (2019).
8. Yoon, C.-H., A validation study of the torrance tests of creative thinking with a sample of Korean elementary school students. *Thinking Skills and Creativity*, 26, 38-50 (2017).
9. Pusca, D. and Northwood, D.O., Technology-based activities for transformative teaching and learning. *World Trans. on Engng. and Technol. Educ.*, 14, **1**, 77-82 (2016).
10. Considine, H., Nafalski, A. and Nedic, Z., Automatic verification of the remote laboratory NetLab. *World Trans. on Engng. and Technol. Educ.*, 17, **1**, 12-16 (2019).
11. Lv, Z., Shi, G. and Yao, M., Teaching reform of an *Automatic Measuring Technology* course based on the CDIO method. *World Trans. on Engng. and Technol. Educ.*, 12, **3**, 550-553 (2014).
12. Pásztor, A., Molnár, G. and Csapó, B., Technology-based assessment of creativity in educational context: the case of divergent thinking and its relation to mathematical achievement. *Thinking Skills and Creativity*, 18, 32-42 (2015).
13. Xidao, L. Yuxiang, X., Lili, Z., Xin, Z., Chen, L. and Jingmeng, H., An image similarity acceleration detection algorithm based on sparse coding. *Mathematical Problems in Engng.* (2018).