JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 12(1), 91–109. https://doi.org/10.18608/jla.2025.8571

Assessing Creativity Across Multi-Step Intervention Using Generative AI Models

Eran Hadas[1] and Arnon Hershkovitz[2]

Abstract
Creativity is an imperative skill for today's learners, one that has important contributions to issues of inclusion and equity in education. Therefore, assessing creativity is of major importance in educational contexts. However, scoring creativity based on traditional tools suffers from subjectivity and is heavily time- and labour-consuming. This is indeed the case for the commonly used Alternative Uses Test (AUT), in which participants are asked to list as many different uses as possible for a daily object. The test measures divergent thinking (DT), which involves exploring multiple possible solutions in various semantic domains. This study leverages recent advancements in generative AI (GenAI) to automate the AUT scoring process, potentially increasing efficiency and objectivity. Using two validated models, we analyze the dynamics of creativity dimensions in a multi-step intervention aimed at improving creativity by using repeated AUT sessions (N=157 9th-grade students). Our research questions focus on the behavioural patterns of DT dimensions over time, their correlation with the number of practice opportunities, and the influence of response order on creativity scores. The results show improvement in fluency and flexibility, as a function of practice opportunities, as well as various correlations between DT dimensions. By automating the scoring process, this study aims to provide deeper insights into the development of creative skills over time and explore the capabilities of GenAI in educational assessments. Eventually, the use of automatic evaluation can incorporate creativity evaluation in various educational processes at scale.

Notes for Practice

- Generative AI for automated creativity assessment: Generative AI provides an automated solution for assessing creativity, exemplified here through Guilford's Alternative Uses Test (AUT).
- Impact of AUT interventions: An AUT intervention enhances fluency and flexibility over time without compromising originality levels.
- Understanding originality: The behaviour of originality over time is more accurately explained by examining aggregated scores that account for the serial order effect.

Keywords: Creativity, divergent thinking (DT), Alternative Uses Test (AUT), generative AI (GenAI), longitudinal study, automated scoring, educational assessment, practice opportunities

Submitted: 28/07/2024 — **Accepted:** 18/02/2025 — **Published:** 19/03/2025

Corresponding author [1]Email: ehadas@tauex.tau.ac.il Address: School of Education, Tel Aviv University, Ramat Aviv, PO Box 39040, Tel Aviv 6997801, Israel. ORCID iD: https://orcid.org/0009-0005-1531-2087
[2]Email: arnonhe@tauex.tau.ac.il Address: School of Education, Tel Aviv University, Ramat Aviv, PO Box 39040, Tel Aviv 6997801, Israel. ORCID iD: https://orcid.org/0000-0003-1568-2238

1. Introduction

Creativity has long been recognized as a crucial skill for learners, employees in the job market, and citizens navigating societal challenges, as well as for personal development. It is also seen as a means to enhance equity among groups of learners from diverse backgrounds (Kozlowski & Si, 2019; Luria et al., 2017). Importantly, creativity is not a fixed inborn trait, and training programs may improve performance in creativity tests (Israel-Fishelson & Hershkovitz, 2022; Scott et al., 2004). Therefore, many recognize the increasing importance of creativity in education as a key 21st-century skill, with practical implications for understanding how creativity develops over time (Long et al., 2022; Runco, 2008; Thornhill-Miller et al., 2023). Such insights are useful for providing personalized feedback to learners and designing adaptive educational programs that nurture and sustain creative potential effectively.

The Alternative Uses Test (AUT; Guilford, 1967) is a widely employed psychological tool for assessing creativity. During the test, participants are tasked with generating multiple uses for a daily object, like a toothbrush or a disposable cup. The test measures verbal divergent thinking (DT; Guilford, 1967), which involves exploring multiple possible solutions in various

semantic domains. Performance on the AUT is typically measured based on four aspects, referred to as the dimensions of DT (see Guilford, 1967; Torrance, 1969, 1974):

1. **Fluency:** The number of eligible responses
2. **Flexibility:** The number of unique conceptual categories of uses
3. **Originality:** The statistical infrequency of uses
4. **Elaboration:** The degree of detail and richness of the responses

Human rating of verbal DT tests, like the AUT, is extremely labour- and time-consuming, which can explain the limited assessment points taken even in long experiments. For instance, Ritter et al. (2020) used three such points over a course of a year, and Israel-Fishelson and Hershkovitz (2022) scored only two over ten weeks, although they had administered 40 additional AUT sessions in between. Moreover, since human judgment is subjective, it requires multiple raters to perform a single evaluation task. A solution to the limitations of human raters is to automate the process of scoring. However this is a non-trivial task for the AUT flexibility and originality metrics.

Flexibility, a key dimension of divergent thinking that reflects the ability to shift between concepts, is difficult to score automatically due to the open-ended and textual nature of responses, as it requires generating semantic categories that cover the entire range of possible answers (Grajzel et al., 2023; Johnson et al., 2021). Originality, which assesses the novelty of generated ideas, might appear easier to score, as it involves assigning a score to a text with available data. However, prior to the advent of generative AI (GenAI), scoring originality yielded significantly weaker results (Organisciak et al., 2023).

Recent advancements in GenAI not only enable content generation but also facilitate the assessment of human creativity (Acar, 2023). In the context of the AUT scoring process, this involves generating appropriate semantic categories, classifying responses into these categories, and evaluating the originality of each response. Notably, the task of devising categories has no previous parallel, highlighting the unique capabilities of GenAI in this domain. For other tasks, the goal is to enhance performance compared to previous models by leveraging GenAI's contextual understanding. Using GenAI, performance on both flexibility and originality dimensions has greatly improved, yielding results comparable to and correlated with human ratings.

DT is not a fixed trait but a skill that evolves over time with repeated practice and exposure (Scott et al., 2004; Valgeirsdottir & Onarheim, 2017). Detailed assessments empower educators to adapt their classroom strategies, offering targeted feedback and resources that nurture each student's unique creative development (Ezzat et al., 2017; Redifer et al., 2021). Generative AI's advanced scoring capabilities facilitate a longitudinal approach to creativity assessment, capturing nuanced insights often overlooked in short-term pre- and post-test models. Combined with learning analytics, this approach provides a comprehensive view of fluency, flexibility, originality, and elaboration as creativity evolves.

By pinpointing key moments of growth, such as increased originality (Beaty & Silvia, 2012; Gonthier & Besançon, 2024), educators can implement targeted, timely interventions that enhance learning outcomes (de Chantal & Organisciak, 2023; Wise & Kenett, 2024). Automated scoring provides systematic and scalable insights into DT dimensions, informing pedagogical strategies and curricular adjustments. However, AI-generated scores may exhibit biases or overlook cultural nuances, making teacher input essential in ensuring that automated assessments align with diverse educational contexts (Hickman et al., 2024; Sporrong et al., 2024).

Therefore, data-driven methods for assessing creativity support educational equity by making this task more scalable and accessible, even in resource-limited settings, which in turn can shape inclusive curricula that cater to diverse learning needs. To capture the full impact of such interventions, it is crucial to evaluate creativity across multiple dimensions: progress over time, performance in specific tasks of varying difficulty, and individual students' thought processes. Examining these facets provides a comprehensive view of creativity's development, enabling tailored interventions for individual and group needs.

This study applies two validated models (Hadas & Hershkovitz, 2024; Organisciak et al., 2023, 2025) to explore creativity dynamics over time within a longitudinal training context using repeated AUT sessions. The detailed outputs of the models allow multi-level learning analytics across different tests, student performance within the same test, and individual responses to objects. Therefore, our research questions are the following:

RQ 1: How do fluency, flexibility, originality, and elaboration behave over time during an AUT-based intervention?

RQ 2: How are fluency, flexibility, originality, and elaboration correlated with a student's number of practice opportunities?

RQ 3: How are fluency, flexibility, originality, and elaboration correlated with the order of uses given within a student's response to an object?

2. Related Work

2.1. Nurturing Creativity

Divergent thinking (DT) is a process that involves the generation of multiple diverse ideas to solve a concrete or abstract problem (Guilford, 1967). DT tests are popular techniques for measuring creativity (Plucker et al., 2011). As training can improve creativity, much research has been done to create such programs, which was extensively reviewed in Scott et al.

(2004) and later in Valgeirsdottir & Onarheim (2017). These reviews show that while there exists a range of well-documented and successful creativity training programs, it is difficult to compare them to determine which are most effective. They range from cognitive, personal, and social models to even building computer-based idea maps (M. Sun et al., 2019).

Interestingly, even relatively simple interventions have shown to be effective. For example, even short 30-minute meditation sessions within a week improved DT in undergraduate students (Ding et al., 2014), and eight 20-minute alternative-uses-generation practice sessions within two weeks improved DT for both adults and adolescents (C. E. Stevenson et al., 2014). Even a single 1.5-hour creativity training session improved undergraduate university students' AUT flexibility (Ritter & Mostert, 2017).

Longer interventions were studied to further understand the effect of creativity training on various DT dimensions, with varied outcomes. A two-semester creativity training, involving 140 hours of related theory and practice was conducted on university students. Four tools — namely simplify, differentiate, visualize, and tag the problem — were tested over various types of creativity steps, including DT, and improvement was reported in fluency and flexibility, but no significant improvement was recorded in measuring originality (Ritter et al., 2020). Other examples include six DT task intervention sessions of 15–25 minutes each over two weeks with university students, who showed some improvement in originality, yet flexibility remained stable over time, and sometimes even decreased (Fahoum et al., 2023). On the other hand, an 8-session training at home for adolescents improved fluency, yet originality did not significantly change (Cousijn et al., 2014).

While AUT serves both as a training activity and an assessment tool, it is not common to see AUT results measured for each individual session, despite the repetitive nature of such interventions. This is mostly due to the difficulty of scoring this tool. However, in some cases, it has been done because pre- and post-test evaluation did not suffice. For instance, to measure differences between a group of adults and a group of adolescents, AUT interventions were performed and scored in the in-between sessions, in addition to pre- and post-tests, where participants were asked to train eight times with a minimum of one day and a maximum of two days between training sessions within two weeks (C. E. Stevenson et al., 2014). AUT multi-step scoring was also used to analyze brain activity: a correlation between university students' brain activity and their scores on AUT after undergoing cognitive simulation training intervention was demonstrated, indicating an increase in both fluency and originality throughout the 20 daily sessions (J. Sun et al., 2016). When the intervention was given to two different groups at different times, there was a pre-test, a test after intervention for group 1 but before intervention for group 2, and a post-test (van de Kamp et al., 2015). In each of these examples, human raters performed the scoring.

2.2. Evaluating Creativity

This study utilizes two novel models to evaluate DT, specifically targeting the dimensions of AUT. To evaluate the originality dimension, we use the model described by Organisciak et al. (2023); to evaluate the flexibility dimension we use the model outlined in Hadas & Hershkovitz (2024). Although both models leverage GenAI to perform analogous tasks, each uses a different approach that continues a separate line of study.

The motivation behind using automatic scoring for DT tests is to replace human scoring, which is prone to subjectivity. Achieving agreement among human raters in DT creativity assessment is a difficult task requiring expertise; it is difficult to explicitly reason, and the standards upon which raters rely are not explicitly articulated (Reiter-Palmon et al., 2019; Silvia et al., 2008; Zedelius et al., 2019), thus leading to errors and biases (Forthmann et al., 2017; Leckie & Baird, 2011; Wilson & Case, 2000). Moreover, human scoring has a high labour cost (Beaty & Johnson, 2021).

Therefore, attempts to automatically score verbal DT tests began long before GenAI, at least in the early 1970s, by finding keywords within responses (Forthmann & Doebler, 2022; Paulus et al., 1970), followed by algorithms calculating the semantic distance between words (Latent Semantic Analysis and Word Embeddings) in order to distinguish between different types of semantic responses to the test (Beaty & Johnson, 2021; Dumas et al., 2021; Olson et al., 2021). The use of large language models (LLM) for assessing creativity is only in its infancy; however, by using word embeddings and pretrained models, it has shown improvements compared to earlier models (Beaty & Johnson, 2021; Patterson et al., 2023; C. Stevenson et al., 2022; Yu et al., 2023). Organisciak et al. (2023) developed Open Creativity Scoring to provide various computational methods to assess creativity, and while starting with LLM-based semantic distance, they introduced a generative-AI method, Open Creativity Scoring with Artificial Intelligence (Oscai; Organisciak et al., 2023), based on prompting ChatGPT to evaluate the dimension of originality. This method, which has shown better performance than older methods, has already been used in other studies (Wahbeh et al., 2024).

In recent years, a new line of research has emerged that seeks to leverage GenAI to boost creativity. GenAI has been demonstrated to match, or even surpass, human creativity in creativity tests by generating innovative responses (Haase & Hanel, 2023; Hubert et al., 2024). Additionally, in tasks such as creative coding and various design challenges, GenAI models have proven to be potent tools, offering valuable insights and support in DT tasks (Acar, 2023; Hwang, 2022; Wingström et al., 2024). Hadas and Hershkovitz (2024) are adopting such an approach to assess the flexibility dimension, which involves understanding the number of unique conceptual categories for suggested uses. This challenge resembles a design problem that the model addresses by dividing into a sequence of two sub-tasks: generating these categories and then categorizing the

responses. The generation of categories in this case is a task that has no parallel in previous automatic approaches, which may have the capacity to split the responses into groups, but not to provide them with meaningful names.

2.3. Usage of AUT Scoring Models

Ocsai is a fine-tuned LLM used for scoring originality in DT tasks, with a primary focus on AUT (Organisciak et al., 2025, 2024). It significantly outperforms traditional semantic distance methods, achieving correlations between Ocsai's scores and human judgments as high as r = 0.81, compared to r = 0.12–0.26 for earlier models using semantic distance. Since its introduction, Ocsai has been continuously maintained and upgraded to utilize the latest versions of ChatGPT.

The model has evolved from using earlier LLMs in the ChatGPT family. It is accessible to users via an API, where users can upload files containing records of objects, language, and responses. The model then returns a score for each record — a number between 1 and 5. The interface is available through a website (https://openscoring.du.edu/scoringllm), with all interactions with the LLM occurring on the backend, meaning users do not directly interact with the LLM.

Organisciak et al. (2023) present a prompt used in a zero-shot version of the model, where the LLM generates scores without any prior task-specific fine-tuning. The zero-shot prompt is:

"What is the originality of the following use for {object} on a scale of 1 to 5: {response}."

However, one of Ocsai's key strengths is its ability to specialize in AUT scoring through fine-tuning. ChatGPT allows this fine-tuning process, where users can upload a file containing prompts and corresponding scores. In this case, the fine-tuning prompt mirrors the zero-shot prompt, and the corresponding scores are human ratings previously collected, with the model learning to match human scores through training. Once fine-tuned, the model is ready for users, who are not involved in the training process. Users can use the model by uploading a list of records containing the object name, language, and provided response. An example format is:

Newspaper, English, Eye Cover
Newspaper, English, To hit the flies

The output is a csv file containing the given object, the given response, and the calculated originality score. For example:
newspaper, eye cover, 3
newspaper, to hit the flies, 2.5

To calculate AUT creativity scores, we adopt the approach suggested by Hadas and Hershkovitz (2024). This approach relies on an automated workflow that utilizes the ChatGPT API, with prompts being issued by a computer program rather than through a website interface. The workflow implements a multi-turn conversation with the LLM, meaning that it introduces a prompt, receives a response, and then generates a subsequent prompt based on that response. In the first phase, the model asks for distinct categories for the responses to an object: "Please examine the responses and determine the distinct categories into which you would assign the responses." In the second phase, it requests the classification of individual responses into the generated categories: "For each response, please give me its most relevant category."

In practice, the prompt is more detailed, as it forces a specific output format and evaluates the effectiveness of the number of categories generated. For several categories close to the square root of the total responses, the model achieved a correlation with human raters of approximately r = .9, as opposed to r < .25 in semantic distance-based attempts (Hass, 2017). The first phase leverages the LLM to overcome the limitations of older methods by generating distinct, semantically meaningful categories based on the responses in the AUT. Crucially, these categories have meaningful names, unlike previous clustering methods, which merely grouped responses without semantic relevance. The second phase improves on earlier techniques, such as semantic distance calculations or topic modeling (Chan & Schunn, 2015; Hass, 2017), by assigning each response to the most appropriate category while taking the AUT context into account. This contextual awareness is essential for calculating flexibility scores.

For example, for the object "shoe," the generated categories could include "Art and Craft Supplies," "Animal-related Uses," and others. When classifying the responses for the shoe test, the output is a CSV file containing the responses and their assigned categories, such as:

mouse trap, Animal-related Uses
painting surface, Art and Craft Supplies

2.4. Response Level Aggregation

While fluency and flexibility are calculated based on the complete set of responses, originality and elaboration are assessed

individually for each student's response to an AUT object. Regarding originality, this allows for various options to provide an aggregated score for the student, representing a major design decision in the scoring procedure (Reiter-Palmon et al., 2019). A significant observation in this field is known as the Serial Order Effect. Due to the constrained time available, initial ideas tend to be relatively mundane. As the test progresses, students typically shift towards generating more original ideas (Bai et al., 2021; Beaty & Silvia, 2012). While later responses tend to be more original, evidence suggests that students who generate fewer ideas overall often produce more original and elaborate responses. This observation indicates the potential benefit of alternative scoring methods for originality, such as using the "best-N ideas" approach, implementing threshold counts, comparing ideas generated in the first half versus the second half of the sequence, or combining them (Gonthier & Besançon, 2024; Wang et al., 2017). We note that currently, elaboration is being computed in a straightforward word-count manner. However, in future studies, if it is computed using GenAI, a similar discussion regarding aggregation may also apply.

3. Methodology

3.1. Participants and Data Collection

The study utilizes data collected in Israel-Fishelson and Hershkovitz (2022). The sample comprised 157 ninth-grade students, 14–15 years old. The experiment sessions, which were part of a broader study, took place as part of the students' school schedule, and were held remotely. Everyday items (e.g., paperclip, ruler, disposable cup) were separately presented to participants, using both a verbal description and an image. Students were given three minutes to think of as many uses as possible for each item and type their responses into an online form. Overall, there were 10 weekly sessions, with four objects presented in each session, i.e., a total of 40 objects; however, attendance and participation were not obligatory.

In this analysis, we will refer to the first objects presented in each weekly session in this order: disposable cup, paperclip, newspaper, key, spoon, shoe, bottle, screwdriver, toothbrush, and cork stopper. A description of our dataset is given in Table **1**. Since our data is based on responses in Hebrew, we first translated them into English; the entire study deals only with the data in English. We used Google Translate and validated it against human translation for the disposable cup. We validated this translation by comparing the automatic flexibility scores for both human-translation and machine-translation, since this dimension is derived from the generated categories, and found the results were strongly correlated: $r(117) = .79$, $p < .001$.

Table 1. Analyzed Objects

Students by Practice Opps		1	2	3	4	5	6	7	8	9	10	Student Total	Responses
Week	Object												
1	Disposable Cup	119										119	446
2	Paperclip	22	88									110	331
3	Newspaper	6	21	64								91	387
4	Key	4	8	18	51							81	237
5	Spoon	5	4	15	13	43						80	274
6	Shoe		3	3	15	11	31					63	204
7	Bottle	2	2	2	6	16	16	25				69	273
8	Screwdriver	1	1	1	2	4	11	11	20			51	180
9	Toothbrush		2	1	5	5	6	9	13	19		60	204
10	Cork Stopper		2	2		1	1	6	8	9	16	45	162

3.2. Model Outline

The first step of the study involves computing the four dimensions of creativity. We use two GenAI models to assess flexibility and originality, while fluency and elaboration are measured using straightforward counting methods. These computations are performed for each student across all tests. Notably, originality is also evaluated independently for each response a student provides for any given object. The data is then analyzed from three distinct perspectives, each representing a different level of detail:

Overall Creativity Trends: A comprehensive repeated measures analysis conducted using a mixed-effects model to explore how creativity metrics evolve over time.

Comparative Performance Analysis: For each object, a separate linear regression model compared student performance within the same test, focusing on their practice opportunities so far.

Detailed Response Analysis: An analysis of the originality of each response within the list provided by a student for an object is conducted, testing the serial order effect.

The methodological framework of the study is outlined in Figure 1.

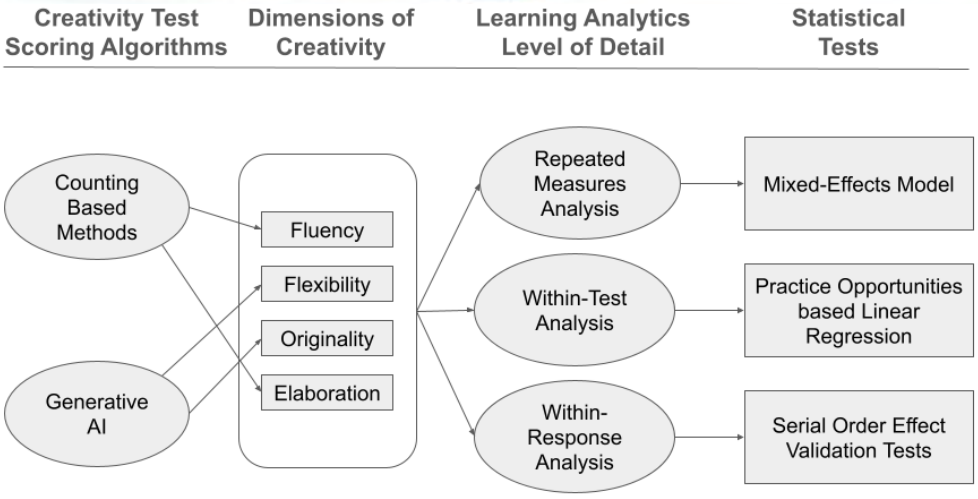

Figure 1. Model outline.

Two recent models utilize the development of user-aligned LLMs, doing so by having the model reply within a conversational environment to a prompt. For calculating originality scores, we use Oscai (Organisciak et al., 2023). The model version used, ocsai-chatgpt, is a GPT-3.5-size chat-based model, trained with the same format and data as previous models, such as GPT-3 Davinci. The model assigns scores on a 1–5 scale, indicating the level of originality of a response, with 1 being very unoriginal and 5 very original. It has shown significant improvements over semantic distance scoring, exhibiting a correlation of r = .813 with human judgments on originality, markedly higher than the r = .256 achieved by the previous semantic distance system.

For flexibility scoring, we employ the approach outlined by Hadas and Hershkovitz (2024). Their method utilizes OpenAI's ChatGPT (GPT-3.5 in the original paper, GPT-4 in our study) without fine-tuning. Specifically, ChatGPT is prompted to generate semantic categories for AUT responses, followed by a second step where each response is assigned to the most relevant category. Their paper outlines various validation methods, adhering to Cohen et al.'s (1996) validity framework, which refers to four categories of validation — face validity, content validity, criterion-related validity, and construct validity — as well as to the detection of biases. It shows a correlation of r = .86 with human judgments on flexibility. The output format is a csv file containing the responses and their determined categories for each object. Despite their promising accuracy, the authors of both methods acknowledge that their approaches are still in their initial stages, necessitating further exploration to fully understand their potential and limitations.

Our study automates the process of scoring. By getting student responses for each object as input, and the mapping tables generated by the GenAI modes, the program loops over each object, scores both the straightforward and the LLM-assisted dimensions by aggregating the scores for each student, and outputs a table with the four scores given for each student for each object. This allows us to add data for more objects, but also for more dimensions, and is performed to evaluate alternate scoring aggregations for the originality dimension.

3.3. Research Variables

Regarding the four dimensions of creativity, automatic scores for fluency, flexibility, originality, and elaboration are computed for each student for each item. As the variables are on different scales, we employ z score normalization for each of them:

Fluency: The number of responses given by a student, computed in a straightforward manner by counting.

Flexibility: We use the ChatGPT 4-based model taken from Hadas and Hershkovitz (2024). The model assigns categories to student responses for each single item. From this, flexibility is computed by counting the distinct resulting mapped categories.

Originality: We use the model "ocsai-chatgpt" taken from Organisciak et al. (2023). The model assigns a score ranging from 1 to 5, where 1 is minimally original, and 5 is maximally original.

Elaboration: The average word-count of student responses, computed in a straightforward manner.

In addition to the four dimensions of creativity, we introduce four more research variables used for analysis among students within the same test, and between different uses for an object provided by individual students. In particular, the alternative scoring aggregation approaches for originality are inspired by Gonthier and Besançon (2024).

Practice Opportunities: For each student, we count the number of AUT tests they have participated in up to and including the current test. Participation was not mandatory, resulting in some students choosing to engage in only a selection of the tests. This approach allows us to describe scores for a single test as a function of practice opportunities, enabling a comparison of student performance within the same test:

Originality Best Two: Defined as the average between the two highest scores on the list. Considering an individual student response for an AUT object, it consists of a list of uses for that object.

Originality Top Index: Defined as the use number that got the highest score on the uses list. In case of equal scores, the lower index is considered. Considering a single response for an AUT object, it consists of a list of uses for that object.

Originality Diff Means: Defined as the difference between the mean originality score of the latter half of their responses and that of the first half.

We also examined two similar variables, Normalized Originality Top Index, which is the Originality Top Index divided by fluency, as well as Originality Top Two Indices, which is the average of the indices giving the best two original scored responses. We briefly report on them when they differ from their equivalents.

For example, if a student listed four uses — scoring 1.5, 2.5, 2.0, 1.0 respectively — the student's Originality Best Two score for this object is 2.25, averaging the second and third responses, which got the highest scores. Originality Top Index for the student for this object is 2, as the second use got the highest score. Originality Diff Means for this the student for this object is -.5, the difference between 1.5 and 2.0. Normalized Originality Top Index is .5, which is 2 divided by 4. Originality Top Two Indices is 2.5, averaging indices 2 and 3.

4. Results

4.1. DT Dimensions Behaviour Over Time (RQ1)

We report on the findings based on the first research question, examining how fluency, flexibility, originality, and elaboration behave over time during an AUT-based intervention. The analysis is implemented using a Python program, utilizing the pandas and numpy libraries for the calculations and matplotlib for plotting the graphs.

To provide an initial, intuitive exploration of the AI-generated scores, we analyzed the distributions of the four DT dimensions across the training program, N = 769 instances. This descriptive analysis aims to reveal the scoring tendencies of the AI model and offers insight into its behavioural patterns. Table 2 presents statistical comparisons for the four DT dimensions.

Table 2. Descriptive Statistics of DT Dimensions Across the Study

	Fluency	Flexibility	Originality	Elaboration
avg	5.16	3.59	2.63	2.98
std	4.22	2.15	0.58	1.99
median	4.00	3.00	2.70	2.50
Q1	2.00	2.00	2.27	1.83
Q3	7.00	5.00	3.00	3.60
IQR	5.00	3.00	0.73	1.77
Skew	2.09	0.77	-0.34	3.09
Kurt	7.35	0.05	0.77	16.38

While fluency and elaboration are calculated in a straightforward manner, flexibility and originality are scored by the LLM-based models. Flexibility scores exhibit a moderately right-skewed distribution, with a mean of M = 3.59 (SD = 2.15), a median of 3.00, and a positive skewness of .77, indicating that higher scores extended the tail to the right. The kurtosis value near zero (.05) suggests a mesokurtic distribution without significant outliers. A wide interquartile range (IQR = 3.00), from Q1 = 2.00 to Q3 = 5.00, reflects substantial variability within the middle 50% of scores. This indicates that the AI model assigns flexibility scores that are diverse and span a broad range, with a moderate tendency towards higher values but without excessive clustering or extreme deviations.

Originality scores exhibit a slightly left-skewed distribution, with a mean of M = 2.63 (SD = 0.58) and a median of 2.70, indicating that scores centre below the scale's mid-point (3 on a 1–5 scale). The negative skewness (skewness = –0.34) suggests a minor skew towards higher scores, with fewer low scores extending the tail to the left. Despite this skew, both the mean and median being below 3 indicate that the AI model tended to assign originality scores below the mid-range. The moderate kurtosis (0.77) implies that the distribution is somewhat peaked around the mean, showing that scores are concentrated near the centre with fewer extreme values. The interquartile range (IQR = 0.73), from Q1 = 2.27 to Q3 = 3.00 further confirms that the scores cluster around, but slightly below, the mid-point. However, the negative skew and compressed variability suggest a potential form of score inflation, which is commonly observed in LLM-based assessments (Hickman et al., 2024). To further investigate this, alternative scoring aggregations are examined to assess potential inflation effects and distributional biases in originality scores (see 2.4).

Fluency and flexibility scores display distinct patterns in their distributions. Fluency scores show a high mean (M = 5.16), and a larger standard deviation (SD = 4.22) compared to flexibility. The high skewness (2.09) and kurtosis (7.35) in fluency imply a strong right skew with a peaked distribution, where many scores cluster at lower values but with some high outliers extending the distribution. This may suggest that students who gave many responses had repeated the same categories.

Elaboration scores present a right-skewed distribution with a mean of M = 2.98 (SD = 1.99) and a median of 2.50, reflecting a central tendency near the lower end of the scoring range. The interquartile range (IQR = 1.77), from Q1 = 1.83 to Q3 = 3.60, shows moderate variability, and the high kurtosis (16.38) suggests a heavily peaked distribution with extreme values, meaning that scores are often concentrated around lower values with some significant outliers.

The correlation matrix for the four dimensions is shown in Table 3. There is a very strong positive correlation (r = .89) between flexibility and fluency. This correlation is significant at the $p < .001$ level. Elaboration has negative correlations with the other three variables, suggesting that lower elaboration scores might be associated with higher scores in flexibility and fluency. However, the relationship with originality is negligible.

Table 3. Correlation Matrix for Dimensions of DT

	Fluency	Flexibility	Originality	Elaboration
Fluency	1.00	**0.89**	0.12	-0.34
Flexibility	**0.89**	1.00	0.34	-0.25
Originality	0.12	0.34	1.00	0.00
Elaboration	-0.34	-0.25	0.00	1.00

To analyze the repeated measures from the AUT sessions, we compared the mean scores given by the model to each dimension in 10 tests, as outlined in Figure . In addition, we employed a mixed linear model (MixedLM) with each dimension's normalized grade as the dependent variable. This mixed-effects model, a standard approach for analyzing repeated measures, was chosen to capture random effects — such as changes over time — that capture the variance between students. Additionally, it provides robust estimates in the presence of missing data, as many students completed only some of the AUT sessions. The assumption is that the scores are affected by the difficulty of the AUT object test, which is expected to be expressed as an effect of the week number. Therefore, the scores are normalized for each session. In addition to the week number variable, we aim to test whether the number of practice opportunities affects the scores. For example, we want to determine if a student performs better in their third session compared to their first session. Therefore, we consider only students who participated in at least two sessions.

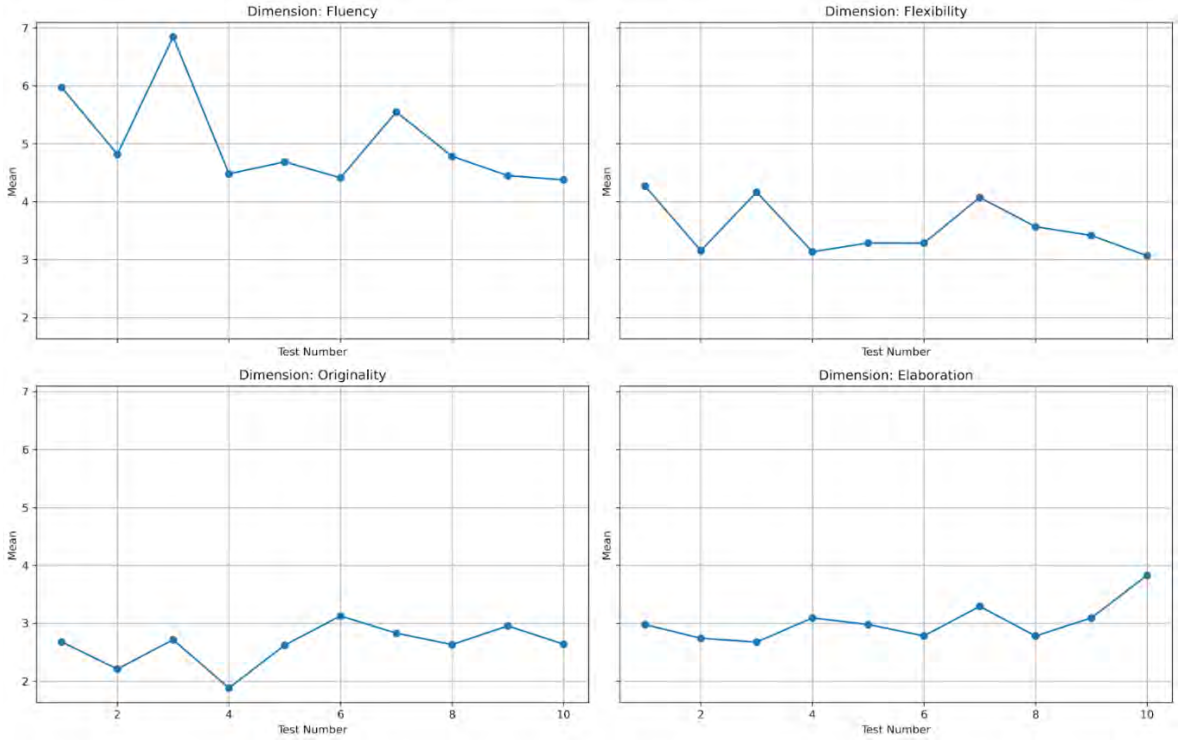

Figure 2. DT dimension means over time.

The models converged for all four dimensions. The analysis included 741 observations, using the REML method. The dataset comprised 131 groups, with a minimum group size of 2, a maximum of 10, and an average group size of 5.7. All models included random intercepts and slopes for students across weeks, allowing for individual differences in both initial performance (intercepts) and change over time (slopes). There is a significant negative effect of week number on fluency scores (β = -.09, p < .01). and on flexibility scores (β = -.07, p < .05). This suggests that as the weeks progress, fluency and flexibility scores tend to decrease. The number of practice opportunities has a significant positive effect on fluency scores (β = .10, p < .01). This indicates that students who take part in more AUT practices tend to have higher fluency scores. As for flexibility, the number of practice opportunities has a marginally significant positive effect on flexibility scores (β = .07, p < .1). No significant effect was found for either variable on originality or elaboration. The variance of the random intercepts (σ^2_intercept) ranged from .29 to .65 across the four creativity measures. The variance of the random slopes for the week number (σ^2_slope) was smaller, ranging from .00 to .01. The covariances between the random intercepts and slopes ranged from -.01 to -.06.

To assert that the effects are consistent after a significant amount of time, we repeated the analysis for students who took part in at least half of the practice opportunities, which was five sessions. The models converged for all four dimensions. The analysis included 601 observations, using the REML method. The dataset comprised 80 groups, with a minimum group size of 5, a maximum of 10, and an average group size of 7.5. There is a significant negative effect of week number on fluency scores (β = -.15, p < .01) and on flexibility scores (β = -.12, p = .01). The number of practice opportunities has a significant positive effect on fluency scores (β = .16, p < .01) and on flexibility scores (β = .11, p < .05). So, the analysis shows the same trends as the previous one, but the effect of practice opportunities on flexibility becomes statistically significant. No significant effect was found for either variable on originality or elaboration. The variance of the random intercepts (σ^2_intercept) ranged from .21 to .84 across the four creativity measures. The variance of the random slopes for the week number (σ^2_slope) ranged from .00 to .01. The covariances between the random intercepts and slopes ranged from .00 to -.07.

4.2. AUT Scores by Practice Opportunities (RQ2)

To complement the insights provided by mixed-effects modelling, this study incorporates linear regression analyses for each creativity dimension. In these analyses, the number of practice opportunities is treated as the independent variable, and the respective AUT test scores serve as the dependent variable. Linear regression is particularly useful here as it allows for a more focused comparison of student performance on identical AUT tests, without the complexity of tracking changes over time or dealing with repeated measures. Additionally, since there is no missing data in these analyses and we are not modelling individual-level variability, linear regression provides a straightforward and valid approach. This method effectively neutralizes potential variations in test difficulty, ensuring that the relationship between practice opportunities and creativity scores is examined directly and clearly. The trendlines, calculated through linear regression, are illustrated in Figure . Bold font indicates statistical significance (p < .05), while italics denotes marginal statistical significance (p < .1) with respect to the null hypothesis that the slope is equal to zero. The slopes (coefficients) of these trendlines are presented in Table 4. In Figure the values are normalized for comparison between dimensions.

Table 4. Regression Line Slopes for AUT Scores w.r.t Practice Opportunities

Week	Fluency		Flexibility		Originality		Elaboration	
	slope	p	slope	p	slope	p	slope	p
2	1.42	.119	.48	.245	.01	.902	.06	.871
3	**2.00**	**.030**	**.77**	**.037**	**.20**	**.000**	.32	.156
4	**1.34**	**.005**	**.76**	**.002**	-.01	.898	*-.65*	*.061*
5	**.70**	**.041**	*.31*	*.074*	-.02	.680	.03	.870
6	**1.01**	**.006**	**.56**	**.015**	-.02	.832	-.15	.407
7	*.66*	*.077*	**.42**	**.043**	.05	.108	.09	.668
8	.41	.200	*.31*	*.098*	-.02	.568	.12	.427
9	.35	.135	.09	.560	.04	.227	-.02	.912
10	**.62**	**.006**	**.35**	**.012**	.01	.805	-.12	.375

Note: Values in bold are statistically significant; those in italics are marginally statistically significant.

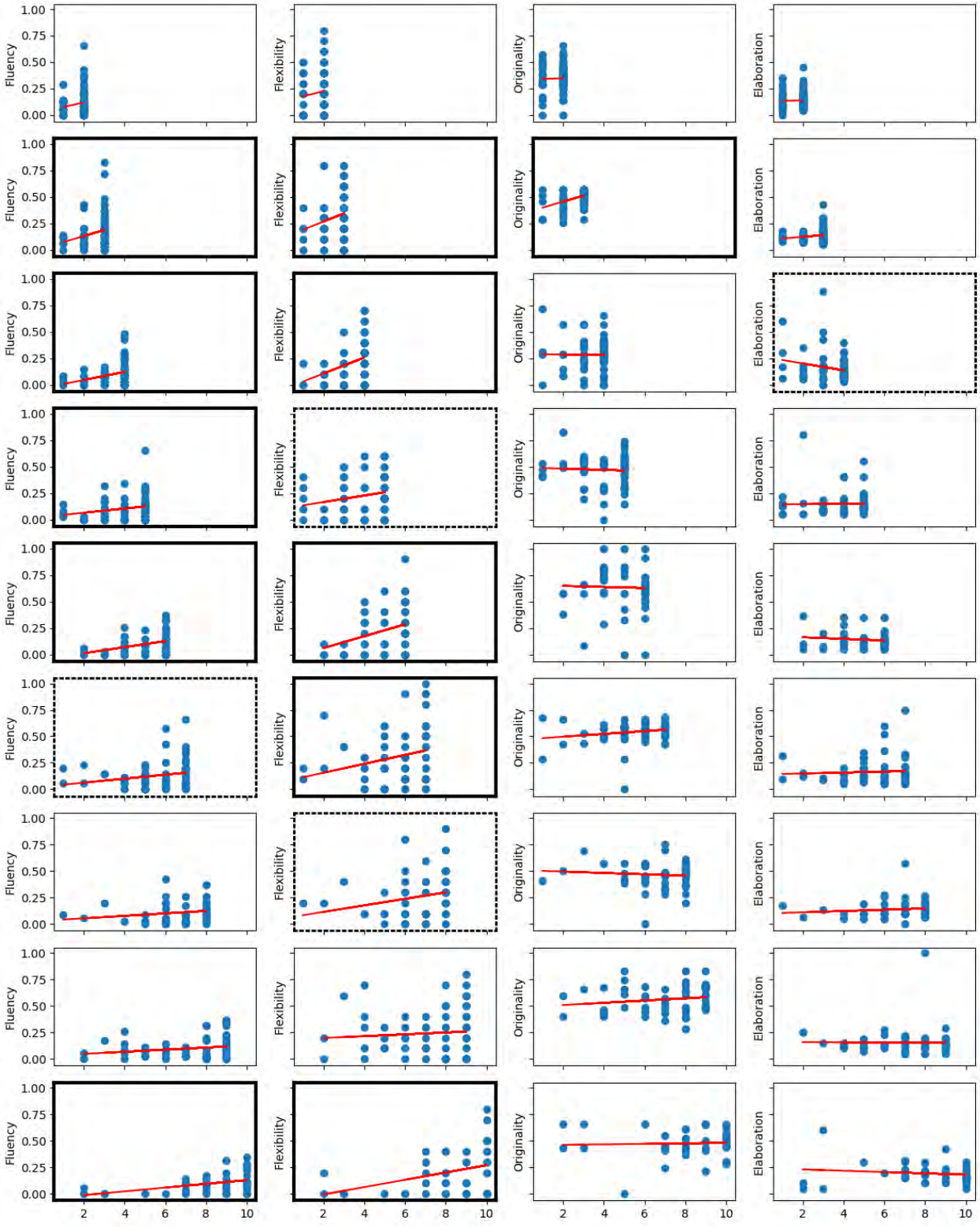

Figure 3. Association between practice opportunities and AUT scores — **normalized** metrics compared (Week 2–10 from top to bottom).

Fluency Scores: In five out of nine tests (tests 3–6 and 10), we find a statistically significant increase in fluency with more practice opportunities, where slope values β range between .62–2.00; in another case (test 6), we find a marginally significant increase, with a slope value of $\beta = .66$, at $p < .1$.

Flexibility Scores: In five out of nine tests (tests 3–4, 6–7, and 10), we found a statistically significant increase in flexibility with more practice opportunities, where slope values β range between .35–.77; in two other cases (tests 5 and 8), we find a marginally significant increase, with a slope value β = .31, at $p < .1$.

Originality Scores: In one case (test 3) we found a statistically significant increase in originality with more practice opportunities, with a slope value of β = .20.

Elaboration Scores: In one case (test 4) we found a marginally significant decrease in elaboration with more practice opportunities, having a slope value of β = .65, at $p < .1$.

4.3. Alternative Scoring Methods for Originality (RQ3)

We examined four different methods for aggregating scores for originality, all using the same scores per response. In addition to the original score, which is the student's average score across responses, we also consider Originality Best Two, Originality Top Index, and Originality Diff Means. We compare the mean scores given by the model to each dimension in 10 tests, as outlined in Figure .

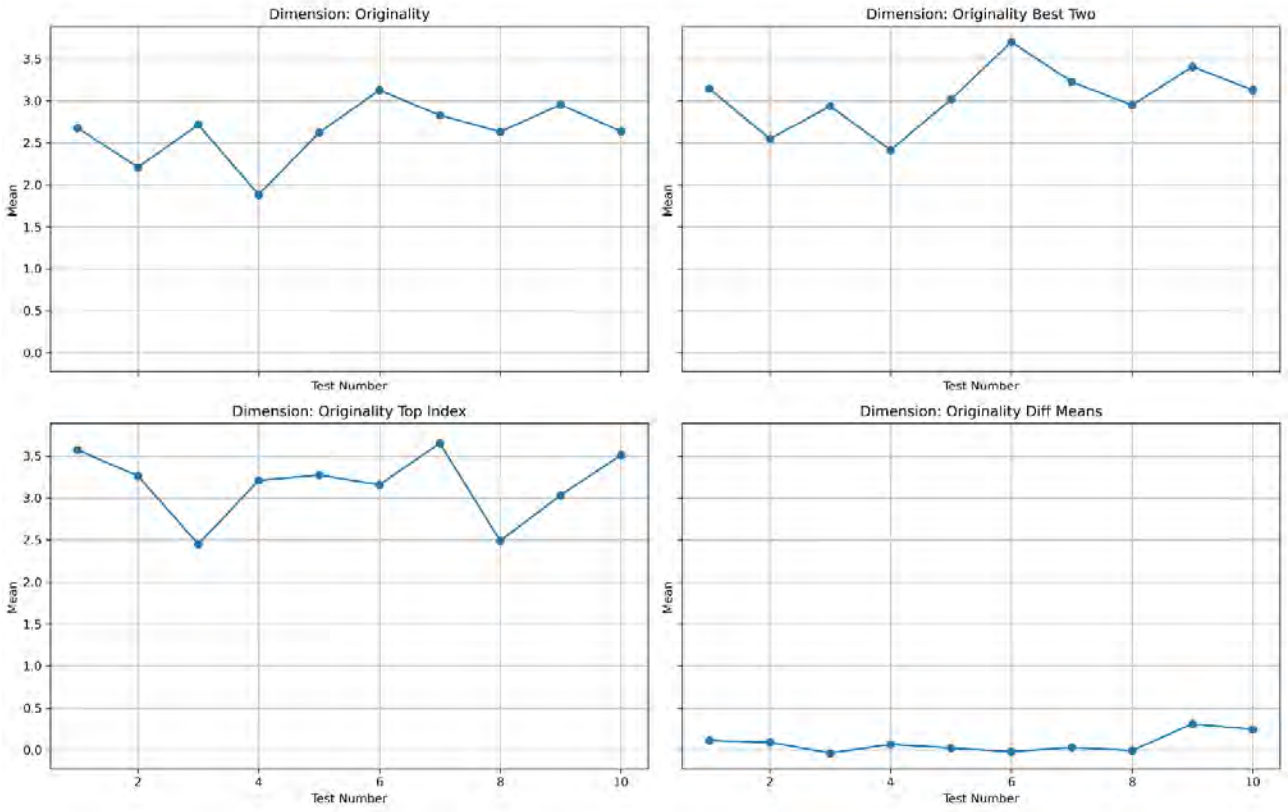

Figure 4. Originality means over time using different aggregations.

We observed a marginally significant positive correlation between Originality Diff Means and elaboration (r = .62, p < .1), as well as between Originality Top Index and elaboration (r = .60, p < .1), indicating that as elaboration increases, the difference in means of originality also tends to increase, and the most original response of the set occurs later. We also tested the case for Originality Top Two indices, the average between the indices of the two most original responses; the correlation with elaboration is even stronger than for Originality Top Index (r = .73, p < .05).

Furthermore, when we normalize the Originality Top Index by dividing it by fluency (the number of responses), this normalized index shows a significant positive correlation with elaboration (r = .68, p < .05). It also exhibits a negative correlation with fluency (r = -.82, p < .01) and flexibility (r = -.71, p < .05). Originality Best Two shows a very high correlation with originality (r = .96, p < .01).

We employed a mixed linear model (MixedLM) with each dimension's normalized grade as the dependent variable. The models converged for all four dimensions. The analysis included 741 observations, using the REML method. The dataset comprised 131 groups, with a minimum group size of 2, a maximum of 10, and an average group size of 5.7. For Originality Top Index and Originality Diff Means, the tests converged but yielded no statistically significant results.

There is a significant negative effect of week number on Originality Top Two Indices scores (β = -.09, p < .01) and on Originality Best Two scores (β = -.07, p < .05), but a significant positive effect on Originality Top Index Normalized (β = .07, p < .05). Practice opportunities has a significant positive effect on Originality Top Two indices scores (β = .12, p < .01), a

marginally significant positive effect on Originality Best Two scores (β = .08, p < .1), and a significant positive effect on Originality Top Index Normalized (β = -.09, p < .05).

We also investigated the four aggregations as a function of practice opportunities for each week separately. The trendlines, calculated through linear regression and normalized, are illustrated in Figure . Bold frames indicate statistical significance (p < .05), while dashed frames denote marginal statistical significance (p < .1) with respect to the null hypothesis that the slope is equal to zero.

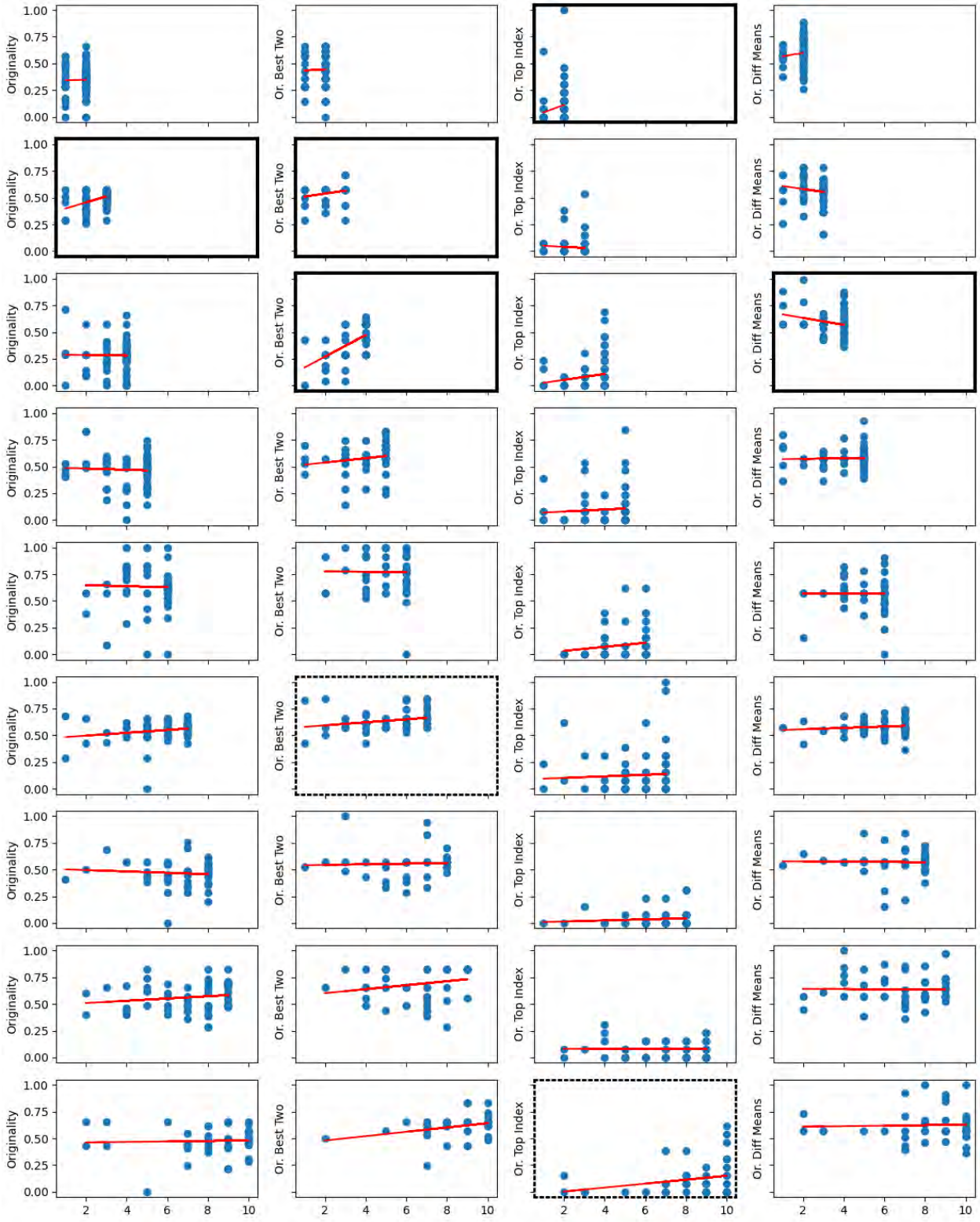

Figure 5. Association between practice opportunities and AUT originality scores — **normalized** metrics compared (Week 2–10 from top to bottom).

Originality Best Two Scores: In two out of nine tests (tests 3 and 4), we find a statistically significant increase in Originality Best Two scores with more practice opportunities, where slope values are .028 and .101, respectively; in another case (test 7), we find a marginally significant increase, with a slope value of $\beta = .014$, at $p < .1$.

Originality Top Index Scores: In one out of nine tests (test 2), we find a statistically significant increase in Originality Top Index with more practice opportunities, where the slope value is $\beta = .069$; $p < .05$. In another case (test 10), we find a marginally significant increase, with a slope value of $\beta = .018$, at $p < .1$.

Originality Diff Means Scores: In one out of nine tests (test 4), we find a statistically significant decrease in Originality Diff Means with more practice opportunities, where the slope value is $\beta = -.033$; $p < .05$.

To explore the behaviour of originality more thoroughly, we investigated the serial order effect by examining whether responses given in the first half of sessions are less original than those in the second half. We conducted a one-sample T-test for the Originality Diff Means metric each week. In three out of ten tests (tests 1, 9, and 10), the positive difference was statistically significant ($p < .05$), and in another test (test 2), it was marginally significant ($p < .1$). However, it is important to note that the absolute value of the average difference was less than 1 for all weeks.

In addition, the behaviour of Originality Top Index and Originality Top Index Normalized (divided by fluency) is outlined in Figure . The values of Originality Top Index range between 2.45–3.65, which shows that indeed the most original response is generally not the first one. The value of the normalized metric in test 3 is .34 and in test 8 is .47, but in the other 8 tests out of 10, it is more than .5, ranging between .58–.72. This indicates that the most original response is given in the second half of the response list.

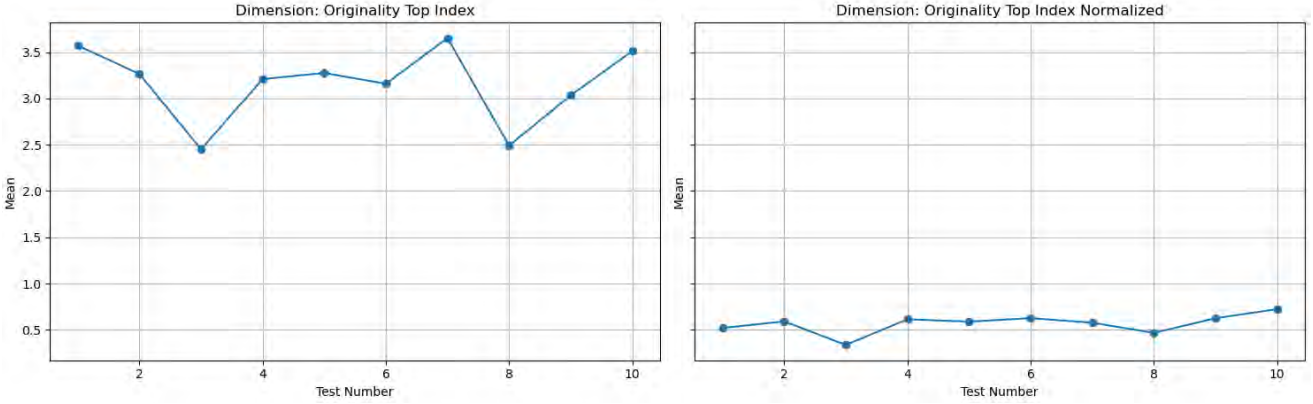

Figure 6. Originality top index means over time.

5. Discussion

5.1. Use of Generative AI for Analysis

Our automatic use of GenAI models allows for a comprehensive scoring of AUT along its four dimensions — namely, fluency, flexibility, originality, and elaboration — in an easy manner. We demonstrated the power of our computational approach to creativity scoring by applying our pipeline to data collected during a 10-week AUT-based intervention program that was previously shown to improve creativity and problem-solving overall from pre-test to post-test (Israel-Fishelson & Hershkovitz, 2022). Each RQ contributes specific understanding: from tracking development (RQ1) and the impact of repeated practice (RQ2) to understanding response patterns (RQ3). This comprehensive approach helps bridge creativity modelling with real-time educational applications. The multi-faceted yet fine-grained analysis enabled by our approach allowed us to have a nuanced understanding of creativity development throughout the program in an unprecedented way. This process of using a previously validated computational model of a construct to study the construct is called Discovery with Models (Baker & Yacef, 2009), and — as we demonstrated here — it can help researchers uncover hidden patterns and relationships, leading to a deeper understanding of various phenomena (Hershkovitz et al., 2013).

With advanced NLP capabilities, LLMs capture subtleties in DT scoring that traditional methods often miss, offering a versatile toolkit for analyzing complex behavioural patterns. Real-time creativity assessment across multiple points allows for consistent tracking, revealing key behaviours or pinpointing changes during specific intervention phases, and supports adaptive educational applications (Acar, 2023). As early use cases suggest (de Chantal & Organisciak, 2023), automated scoring with LLMs provides scalable tools for modelling creativity and fostering targeted interventions in educational, psychological, and social science contexts. Overall, automatic scoring can help with both better understanding of creativity and its nurture.

5.2. Key Findings

5.2.1. Correlations Between DT Dimensions Over Time

The automatic score provided by the GenAI models enables several insights on the behavioural dynamics of DT dimensions over time. There is a very high positive correlation between fluency and flexibility. In addition, while there is no significant correlation between originality and elaboration, when applying different methods of aggregation to originality, some of these methods yield results that exhibit significant correlation with elaboration, namely Originality Top Two Indices and Originality Diff Means. These results provide a binding between each of the GenAI-based scores to a straightforward-computed measure. This could be used as a baseline or a validity test for future automatic scoring methods. The correlation between flexibility and fluency is in line with the literature (Gonthier & Besançon, 2024; Weiss & Wilhelm, 2022), whereas the correlation between originality (alternative aggregation methods) and elaboration can be explained by the dual pathway theory. According to this theory, achieving high originality requires either high flexibility — exploring many idea categories — or persistence, which involves deep thinking focused on a few ideas (Baas et al., 2013; Nijstad et al., 2010). The persistence pathway may or may not be reflected in high elaboration, as the detailed thinking process may be put into writing. More generally, when dual-process models of creativity may apply (Sowden et al., 2018), various scoring approaches may be used in parallel to capture the various aspects of the model (Reiter-Palmon et al., 2019). The impact of practice opportunities on fluency and flexibility illustrates GenAI's utility in adaptive models, supporting dynamic, individualized feedback based on observed progress. Such adaptive systems, informed by automated creativity scoring, represent an innovative direction for educational environments.

While the negative correlation between elaboration and both fluency and flexibility is not statistically significant in the correlation analysis of the dimensions, we did find that Originality Top Index is positively correlated with elaboration and negatively correlated with fluency and flexibility. This suggests a quality–quantity trade-off: as students focus on generating more ideas, they may put less effort into detailing them. Even more generally, the trade-off is between fluency and flexibility versus originality and elaboration. Yet this is not seen when comparing them directly, because other effects influence the behaviour of DT dimensions, namely the test difficulty and practice opportunities. This is consistent with the trade-off between flexibility and elaboration shown by Gonthier and Besançon (2024) and the trade-off between fluency and originality exhibited by Atakaya et al. (2024).

5.2.2. The Effect of Training on DT Dimensions

Despite the varying number of sessions completed by each student and unknown difficulty level of each test, the mixed effects model effectively handled the repeated measures, tracing the effect of practising opportunities on student performance. The robustness of the model outlines the effect but also provides justification to further explore this effect in detail. Indeed, when analyzing the effect of practice opportunities on each test separately, our study shows improvement in fluency and flexibility along an AUT-based intervention program; the increase in fluency and flexibility because of training is in line with previous studies (Scott et al., 2004; C. E. Stevenson et al., 2014). In one such study, using various DT training methods that included AUT, the AUT method resulted in a slight originality score decline from pre-test to post-test (C. E. Stevenson et al., 2014). Previously, a decline in originality was found (Levav-Waynberg & Leikin, 2012) and was explained by the fact that more responses make it harder for students to produce rare responses. Moreover, longer thinking time between answers has shown to be a predictor of originality (Acar et al., 2019), and during a limited timeframe, such as in AUT, a growth in fluency makes thinking time between answers shorter on average; hence originality is expected to decline.

Looking at the practice opportunities analysis in our study for the alternate originality aggregations, while originality shows a significant increase only for week 3, Originality Best Two exhibits a significant increase for both weeks 3 and 4. This may be indicative of the effectiveness of the training for originality, at least during the first four weeks. In addition, when looking at the variables that measure the behaviour within the response sequence, Originality Top Index shows a significant increase for week 2, and Originality Diff Means shows a significant decrease for week 4, which may be attributed to the fact the first half of the replies is improving due to the training. While several instances of Originality Diff Means (sometimes split by time instead of half of the responses) are found in the literature and affirm the Serial Order Effect (Johns et al., 2001; Wang et al., 2017), we did not find a depiction of its behaviour in a multi-step study.

As for elaboration, our results show a non-clear trend, which may suggest stability in this dimension. Since the increase in fluency enables less time to detail each response, a substantial decline in elaboration is expected. And yet, our alternate originality aggregation analysis demonstrated a positive correlation between originality measures and elaboration. This implies that the behaviour of elaboration may involve originality. Despite evidence of the Serial Order Effect, originality was significantly higher for participants who generated fewer ideas (Gonthier & Besançon, 2024). This result was attributed to the possibility that original ideas require elaboration, which aligns with the controlled attention theory of creativity (Beaty et al., 2014).

The positive impact of practice opportunities is somewhat counteracted by the negative — albeit negligible — effect of time; specifically, fluency and flexibility decline as the weeks progress. This negligible negative trend may be an artifact of the observed fluctuations in creativity measures, which may point out to that some objects are more challenging for students

than others. For instance, when we applied the mixed linear model to data from weeks 3–7, it failed to converge for fluency, likely due to these fluctuations. That some items in creativity tests can be more difficult than others was already demonstrated before (Gupta et al., 2012; Kozbelt & Serafin, 2009). Therefore, modifying the order of the objects might mitigate the observed decline in performance. Additionally, a decline in student motivation as the intervention progressed could have further influenced the outcomes, as is also evident from the decline in participation. These issues should be further studied.

5.3. Implications

Our study introduces a model that utilizes GenAI-based scoring techniques to analyze the dynamics of creativity during an intervention aimed at nurturing creative skills. This approach provides both theoretical and practical implications. It demonstrates that creativity changes over time, underscoring the need for longitudinal studies to examine the long-term effects of repeated creativity training. Such studies could focus on various dimensions of creativity and design interventions targeted at specific aspects of creative skills.

The models used in our analysis are novel due to their incorporation of GenAI. This innovation prompts further research into the capabilities of GenAI in scoring creative tasks. Within the four traditional dimensions of creativity, our findings suggest that various aspects should be re-evaluated — from different methods of aggregating originality scores to improving the measurement of elaboration, potentially by leveraging GenAI beyond mere word counting.

Beyond automated scoring, practical classroom applications may benefit from integrating teacher expertise. Teachers could play an active role in reviewing AI-generated creativity scores and refining them where necessary, particularly when automated assessments fail to capture context-specific originality. Allowing teachers to supplement or adjust AI-based scores would enhance reliability and ensure that creativity assessments remain adaptable to diverse learning environments. This hybrid approach aligns with previous research advocating for human–AI collaboration in educational assessments and increases educators' trust in AI-based educational technology (Hickman et al., 2024; Nazaretsky et al., 2022; Sporrong et al., 2024; Viberg et al., 2024).

Finally, our study highlights the potential of integrating automated scoring and feedback systems into creative practices. Evaluating their effectiveness compared to traditional methods could transform how creativity is cultivated, whether as part of a structured learning curriculum or in independent settings. Our results underscore the transformative role of GenAI in real-time creativity assessment, with implications for both scalable creativity training and adaptive learning systems. By reliably tracking creativity trends over time, GenAI supports longitudinal research on creative skill development and opens avenues for personalized learning environments responsive to individual creative growth.

5.4. Additional Takeaways

Our findings and process provide a replicable model for enhancing similar testing approaches in educational and psychological contexts, while also highlighting patterns for consideration in related studies. Although LLM-based models can operate as "black boxes," making them initially challenging to interpret, repeated automated use allows us to analyze their overall behaviour more effectively. In addressing RQ1, we gained a deeper understanding of the model's dynamics by examining the distribution of its results and comparing these with our initial assumptions — such as expecting a right-skewed distribution for flexibility and a slight deviation from a normal distribution for originality.

A key takeaway from the Ocsai originality scoring model is that existing domain-specific data, such as previously collected human ratings, can be seamlessly incorporated into LLM-based models. This approach fosters synergy between models, enhancing performance rather than replacing one system with another. Additionally, we have reproduced the results of the zero-shot version by applying the prompts mentioned above on datasets introduced in two key journal articles (Beaty et al., 2018; Beaty & Silvia, 2012). Our results showed slight improvements when using GPT-4, a newer version. The correlations with human scores increased from $r = .57$ in the original paper to $r = .59$, and from $r = .52$ to $r = .59$, respectively. Moreover, this gave us confidence that our approach aligns well with Ocsai's underlying methodology. To be clear, in our study we use the fine-tuned version.

The key takeaway from using the flexibility scoring model is that even for tasks typically viewed as assignment or classification tasks, it can be beneficial to consider how a generative approach may offer advantages. Moreover, generative use of LLMs can be constrained, meaning it doesn't necessarily yield free text outputs. This approach is scalable, as it was designed to handle large volumes of responses, and the integration into an automated workflow within a running program suggests the potential for applying it across multiple tests.

5.5. Limitations

Our study has some limitations, regarding both data and technology. The data for this study was collected from Hebrew-speaking students and was automatically translated to English for processing. In addition, the analysis is based on students from a single country (Israel), which has specific educational, cultural, and linguistic characteristics. Moreover, students were not required to attend all sessions, which results in incomplete data.

As for technology, the analysis is based on two recent GenAI models, which have been validated on limited data. In addition, they rely on ChatGPT, which is a third-party commercial product, built and marketed by OpenAI, and therefore we do not have control of its future behaviour nor its future data. We expect newer technologies to improve the performance of the model.

One challenge in using GenAI for assessment is ensuring transparency in its decision-making process. Although LLMs can be prompted to explain their classification decisions and allow token probability inspection, these techniques may not always scale efficiently in large-scale automated scoring pipelines. Moreover, such explanations may not always align with human interpretability, particularly in the context of educational assessments where clarity in scoring rationale is essential. Although fine-tuning can address certain needs, general-purpose GenAI models may lack the domain-specific sensitivity necessary for accurate creativity assessment across varied cultural and contextual dimensions.

On the other hand, the efficacy of these models also heavily relies on access to extensive, high-quality, and domain-specific training data. In the absence of such data, the accuracy of the scores may be compromised, especially when applied to diverse populations. Consequently, the models may struggle to accurately assess responses that deviate from established patterns within their training sets.

Finally, given that current GenAI models are primarily text-based, they are less suited to assessing creativity in tasks involving non-textual or multimodal elements. Expanding GenAI to handle multimodal inputs would necessitate significant advancements in model architecture and training methodologies, which may remain challenging at present.

6. Conclusion

This study examines the dynamics of AUT dimensions over time by employing two GenAI models to automate the scoring of AUT in large volumes. The resulting data supports analysis at three levels of detail: using a mixed effects model to trace factors influencing creativity over time, employing practice opportunities as a predictive variable, and implementing various aggregation techniques on scoring results to uncover insights previously inaccessible. Our findings illustrate improvements in fluency and flexibility throughout a multi-step training program, identify specific aspects of originality that can be enhanced, and demonstrate that while elaboration remains stable, it correlates with various facets of originality. This approach can pave the way for further creativity studies as more advanced GenAI models emerge and as learning analytics evolve to manage the data's depth and breadth.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Acar, S. (2023). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal*, 1–7. https://doi.org/10.1080/10400419.2023.2271749

Acar, S., Abdulla Alabbasi, A. M., Runco, M. A., & Beketayev, K. (2019). Latency as a predictor of originality in divergent thinking. *Thinking Skills and Creativity*, *33*, Article 100574. https://doi.org/10.1016/j.tsc.2019.100574

Atakaya, M. A., Sak, U., & Ayas, M. B. (2024). A study on psychometric properties of creativity indices. *Creativity Research Journal*, *36*(2), 348–364. https://doi.org/10.1080/10400419.2022.2134550

Baas, M., Roskes, M., Sligte, D., Nijstad, B. A., & De Dreu, C. K. W. (2013). Personality and creativity: The dual pathway to creativity model and a research agenda. *Social and Personality Psychology Compass*, *7*(10), 732–748. https://doi.org/10.1111/spc3.12062

Bai, H., Leseman, P. P. M., Moerbeek, M., Kroesbergen, E. H., & Mulder, H. (2021). Serial order effect in divergent thinking in five- to six-year-olds: Individual differences as related to executive functions. *Journal of Intelligence*, *9*(2), Article 20. https://doi.org/10.3390/jintelligence9020020

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with *SemDis*: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapil, T. R., Kane, M. J., & Silvia, P. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, *115*(5), 1087–1092. https://doi.org/10.1073/pnas.1713532115

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319. https://doi.org/10.1037/a0029171

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197. https://doi.org/10.3758/s13421-014-0428-8

Chan, J., & Schunn, C. D. (2015). The importance of iteration in creative conceptual combination. *Cognition*, *145*, 104–115. https://doi.org/10.1016/j.cognition.2015.08.008

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement* (3rd ed.). Mayfield Publishing Co.

Cousijn, J., Zanolie, K., Munsters, R. J. M., Kleibeuker, S. W., & Crone, E. A. (2014). The relation between resting state connectivity and creativity in adolescents before and after training. *PLoS ONE*, *9*(9), Article e105780. https://doi.org/10.1371/journal.pone.0105780

de Chantal, P.-L., & Organisciak, P. (2023). Automated feedback and creativity: On the role of metacognitive monitoring in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000592

Ding, X., Tang, Y.-Y., Tang, R., & Posner, M. I. (2014). Improving creativity performance by short-term meditation. *Behavioral and Brain Functions*, *10*(1), Article 9. https://doi.org/10.1186/1744-9081-10-9

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(4), 645–663. https://doi.org/10.1037/aca0000319

Ezzat, H., Camarda, A., Cassotti, M., Agogué, M., Houdé, O., Weil, B., & Le Masson, P. (2017). How minimal executive feedback influences creative idea generation. *PLoS ONE*, *12*(6), Article e0180458. https://doi.org/10.1371/journal.pone.0180458

Fahoum, N., Pick, H., & Shamay-Tsoory, S. (2023). The impact of creativity training on inter-group conflict-related emotions. *Journal of Conflict Resolution, 68*(7–8), 1494–1521. https://doi.org/10.1177/00220027231198517

Forthmann, B., & Doebler, P. (2022). Fifty years later and still working: Rediscovering Paulus et al.'s (1970) automated scoring of divergent thinking tests. *Psychology of Aesthetics, Creativity, and the Arts, 19*(1), 63–76. https://doi.org/10.1037/aca0000518

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139. https://doi.org/10.1016/j.tsc.2016.12.005

Gonthier, C., & Besançon, M. (2024). It is not always better to have more ideas: Serial order and the trade-off between fluency and elaboration in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts, 18*(4), 480–492. https://doi.org/10.1037/aca0000485

Grajzel, K., Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2023). Measuring flexibility: A text-mining approach. *Frontiers in Psychology*, *13*, Article 1093343. https://doi.org/10.3389/fpsyg.2022.1093343

Guilford, J. P. (1967). *The nature of human intelligence.* McGraw-Hill.

Gupta, N., Jang, Y., Mednick, S. C., & Huber, D. E. (2012). The road not taken: Creative solutions require avoidance of high-frequency responses. *Psychological Science*, *23*(3), 288–294. https://doi.org/10.1177/0956797611429710

Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3), Article 100066. https://doi.org/10.1016/j.yjoc.2023.100066

Hadas, E., & Hershkovitz, A. (2024). Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity, 52*, Article 101549. https://doi.org/10.1016/j.tsc.2024.101549

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Hershkovitz, A., Baker, R. S. J. d., Gobert, J., Wixon, M., & Pedro, M. S. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*, *57*(10), 1480–1499. https://doi.org/10.1177/0002764213479365

Hickman, L., Dunlop, P. D., & Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, *32*(4), 499–511. https://doi.org/10.1111/ijsa.12479

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), Article 3440. https://doi.org/10.1038/s41598-024-53303-w

Hwang, A. H.-C. (2022). Too late to be creative? AI-empowered tools in creative processes. In S. Barbosa, C. Lampe, C. Appert, & D. A. Shamma (Eds.), *CHI EA '22: CHI Conference on Human Factors in Computing Systems extended abstracts* (Article 38). ACM Press. https://doi.org/10.1145/3491101.3503549

Israel-Fishelson, R., & Hershkovitz, A. (2022). Cultivating creativity improves middle school students' computational thinking skills. *Interactive Learning Environments, 32*(2), 431–446. https://doi.org/10.1080/10494820.2022.2088562

Johns, G. A., Morse, L. W., & Morse, D. T. (2001). An analysis of early vs. later responses on a divergent production task across three time press conditions. *The Journal of Creative Behavior*, *35*(1), 65–72. https://doi.org/10.1002/j.2162-6057.2001.tb01222.x

Johnson, D. R., Cuthbert, A. S., & Tynan, M. E. (2021). The neglect of idea diversity in creative idea generation and evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 125–135. https://doi.org/10.1037/aca0000235

Kozbelt, A., & Serafin, J. (2009). Dynamic evaluation of high- and low-creativity drawings by artist and nonartist raters. *Creativity Research Journal*, *21*(4), 349–360. https://doi.org/10.1080/10400410903297634

Kozlowski, J. S., & Si, S. (2019). Mathematical creativity: A vehicle to foster equity. *Thinking Skills and Creativity*, *33*, Article 100579. https://doi.org/10.1016/j.tsc.2019.100579

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Levav-Waynberg, A., & Leikin, R. (2012). The role of multiple solution tasks in developing knowledge and creativity in geometry. *The Journal of Mathematical Behavior*, *31*(1), 73–90. https://doi.org/10.1016/j.jmathb.2011.11.001

Long, H., Kerr, B. A., Emler, T. E., & Birdnow, M. (2022). A critical review of assessments of creativity in education. *Review of Research in Education*, *46*(1), 288–323. https://doi.org/10.3102/0091732X221084326

Luria, S. R., Sriraman, B., & Kaufman, J. C. (2017). Enhancing equity in the classroom by teaching for mathematical creativity. *ZDM*, *49*(7), 1033–1039. https://doi.org/10.1007/s11858-017-0892-2

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring teachers' trust in AI-based educational technology. In A. F. Wise, R. Martinez-Maldonado, & I. Hilliger (Eds.), *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 56–66). ACM Press. https://doi.org/10.1145/3506860.3506866

Nijstad, B. A., De Dreu, C. K. W., Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, *21*(1), 34–77. https://doi.org/10.1080/10463281003765323

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, *118*(25), Article e2022340118. https://doi.org/10.1073/pnas.2022340118

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, Article 101356. https://doi.org/10.1016/j.tsc.2023.101356

Organisciak, P., Dumas, D., Acar, S., & de Chantal, P.-L. (2025). *Open creativity scoring* [Computer software]. University of Denver. https://openscoring.du.edu/

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., … Beaty, R. E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, *17*(4), 495–507. https://doi.org/10.1037/aca0000618

Paulus, D. H., Renzulli, J. S., & Archambault, F. X., Jr. (1970). *Computer simulation of human ratings of creativity: Final report.* National Center for Educational Research and Development.

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. *The Journal of Creative Behavior*, *45*(1), 1–22. https://doi.org/10.1002/j.2162-6057.2011.tb01081.x

Redifer, J. L., Bae, C. L., & Zhao, Q. (2021). Self-efficacy and performance feedback: Impacts on cognitive load during creative thinking. *Learning and Instruction*, *71*, Article 101395. https://doi.org/10.1016/j.learninstruc.2020.101395

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

Ritter, S. M., Gu, X., Crijns, M., & Biekens, P. (2020). Fostering students' creative thinking skills by means of a one-year creativity training program. *PLOS ONE*, *15*(3), Article e0229773. https://doi.org/10.1371/journal.pone.0229773

Ritter, S. M., & Mostert, N. (2017). Enhancement of creative thinking skills using a cognitive-based creativity training. *Journal of Cognitive Enhancement*, *1*(3), 243–253. https://doi.org/10.1007/s41465-016-0002-3

Runco, M. A. (2008). Creativity and education. *New Horizons in Education*, *56*(1).

Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. *Creativity Research Journal*, *16*(4), 361–388. https://doi.org/10.1080/10400410409534549

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Sowden, P. T., Pringle, A., & Gabora, L. (2018). The shifting sands of creative thinking: Connections to dual-process theory. In K. J. Gilhooly, L. J. Ball, & L. Macchi (Eds.), *Insight and creativity in problem solving* (pp. 40–60). Routledge. https://doi.org/10.4324/9781315144061-3

Sporrong, E., McGrath, C., & Cerratto Pargman, T. (2024). Situating AI in assessment: An exploration of university teachers' valuing practices. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00558-8

Stevenson, C. E., Kleibeuker, S. W., de Dreu, C. K. W., & Crone, E. A. (2014). Training creative cognition: Adolescence as a flexible period for improving creativity. *Frontiers in Human Neuroscience*, *8*, Article 827. https://doi.org/10.3389/fnhum.2014.00827

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's creativity to the (alternative uses) test*. arXiv. https://doi.org/10.48550/arXiv.2206.08932

Sun, J., Chen, Q., Zhang, Q., Li, Y., Li, H., Wei, D., Yang, W., & Qiu, J. (2016). Training your brain to be more creative: Brain functional and structural changes induced by divergent thinking training. *Human Brain Mapping*, *37*(10), 3371–3699. https://doi.org/10.1002/hbm.23246

Sun, M., Wang, M., & Wegerif, R. (2019). Using computer-based cognitive mapping to improve students' divergent thinking for creativity development. *British Journal of Educational Technology*, *50*(5), 2217–2233. https://doi.org/10.1111/bjet.12825

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, *11*(3), Article 54. https://doi.org/10.3390/jintelligence11030054

Torrance, E. P. (1969). *Creativity.* National Education Association.

Torrance, E. P. (1974). *The Torrance tests of creative thinking: Norms-technical manual.* Personal Press.

Valgeirsdottir, D., & Onarheim, B. (2017). Studying creativity training programs: A methodological analysis. *Creativity and Innovation Management*, *26*(4), 430–439. https://doi.org/10.1111/caim.12245

van de Kamp, M.-T., Admiraal, W., van Drie, J., & Rijlaarsdam, G. (2015). Enhancing divergent thinking in visual arts education: Effects of explicit instruction of meta-cognition. *British Journal of Educational Psychology*, *85*(1), 47–58. https://doi.org/10.1111/bjep.12061

Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., Coelho, R., & Kizilcec, R. F. (2024). What explains teachers' trust in AI in education across six countries? *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00433-x

Wahbeh, H., Cannard, C., Yount, G., Delorme, A., & Radin, D. (2024). Creative self-belief responses versus manual and automated alternate use task scoring: A cross-sectional study. *Journal of Creativity*, *34*(3), Article 100088. https://doi.org/10.1016/j.yjoc.2024.100088

Wang, M., Hao, N., Ku, Y., Grabner, R. H., & Fink, A. (2017). Neural correlates of serial order effect in verbal divergent thinking. *Neuropsychologia*, *99*, 92–100. https://doi.org/10.1016/j.neuropsychologia.2017.03.001

Weiss, S., & Wilhelm, O. (2022). Is flexibility more than fluency and originality? *Journal of Intelligence*, *10*(4), Article 96. https://doi.org/10.3390/jintelligence10040096

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 113–134). Ablex Publishing Corporation.

Wingström, R., Hautala, J., & Lundman, R. (2024). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity Research Journal*, *36*(2), 177–193. https://doi.org/10.1080/10400419.2022.2107850

Wise, T. A., & Kenett, Y. N. (2024). Sparking creativity: Encouraging creative idea generation through automatically generated word recommendations. *Behavior Research Methods*, *56*(7), 7939–7962. https://doi.org/10.3758/s13428-024-02463-8

Yu, Y., Beaty, R. E., Forthmann, B., Beeman, M., Cruz, J. H., & Johnson, D. (2023). A MAD method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000573

Zedelius, C. M., Mills, C., & Schooler, J. W. (2019). Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, *51*(2), 879–894. https://doi.org/10.3758/s13428-018-1137-1