Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Validating AI-Based Scoring of Divergent Thinking in Elementary School Children

Eran HADAS[a*], Ben AVITAL-LEV[a] & Arnon HERSHKOVITZ[a]
[a]*School of Education, Tel Aviv University, Israel*
*ehadas@tauex.tau.ac.il

Abstract: Divergent Thinking (DT), a core aspect of creativity, is commonly assessed using Guilford's Alternative Uses Test (AUT). This study examines the validity of an automated scoring approach based on a Large Language Model (LLM), applied to AUT responses from 106 third- and fourth-grade students. We focused on the flexibility and originality dimensions, evaluating the automated scores against human ratings using content and criterion-related validity. The model showed strong correlations with human ratings for flexibility, and outperformed the best-known benchmark in assessing originality, supporting its validity. These findings suggest that LLM-based scoring offers a scalable and objective alternative for DT assessment in elementary educational contexts.

Keywords: Divergent Thinking, Alternative Uses Test, Elementary School, Creativity Assessment, Large Language Models

1. Introduction

Creativity is widely recognized as essential for success in education and beyond, supporting innovation, problem-solving, and adaptability. At the core of creativity lies divergent thinking (DT)—the ability to generate multiple ideas or solutions in response to a given problem or object (Runco & Acar, 2012). DT is commonly assessed using Guilford's Alternative Uses Test (AUT), in which participants are presented with a daily object, and are asked to supply as many uses for this object as possible, in a given time; this test enables evaluation of four key dimensions: fluency, flexibility, originality, and elaboration. These dimensions capture the quantity, variety, uniqueness, and detail of creative responses, respectively (Guilford et al., 1978).

AUT fluency and elaboration are straightforward to compute, by counting the number of responses and the average length of responses, respectively. However, human scoring of AUT flexibility and originality is subjective, lacks explicit articulation and is labor-intensive and costly (Forthmann et al., 2017; Reiter-Palmon et al., 2019). To address these limitations, automated AUT scoring efforts began in 1970 with a PL/I program that identified keywords in responses (Paulus, 1970). More recent approaches use Latent Semantic Analysis and Word Embeddings to measure semantic distances between words, benefiting from advances in LLMs (Beaty & Johnson, 2021; Buczak et al., 2023). Recently, the first LLM-based methods to show strong correlations with human scoring were presented for originality (Organisciak et al., 2023, Ocsai: 2024) and flexibility (Hadas & Hershkovitz, 2024).

Assessing DT in young children is especially important due to the foundational role creativity plays in early cognitive development, with implications for personalized instruction and intervention (Sio & Lortie-Forgues, 2024). The AUT has been adapted for younger populations to explore the development of creativity from an early age (Gubenko & Houssemand, 2022). Research shows that while children often produce a large number of ideas (high fluency), these ideas are not always diverse in nature (low flexibility) or novel (low originality), with many responses falling within the same semantic category or relying on familiar patterns (Runco & Acar, 2012).

Automatic computation of AUT scores has not yet been applied to elementary school students, a critical stage for assessing DT. The automatic models mentioned above (Hadas &

Hershkovitz, 2024; Organisciak et al., 2024), though validated on adults, have seen limited use and lack validation in younger populations—possibly due to children's developing language, which tends to be simpler and less varied than adults'. The originality scoring benchmark, Ocsai, was fine-tuned on adult data and may not generalize well to children. We address these gaps by applying pre-trained LLMs, using prompt-based protocols without fine-tuning, to assess flexibility and originality in young learners—offering a novel benchmark for scalable, age-appropriate creativity assessment.

This study evaluates an LLM-based approach for scoring DT in 106 third- and fourth-grade students using an AUT prompt for the object chair. Flexibility and originality were assessed through ChatGPT 4o-generated scores. The model's outputs were compared to human ratings to examine reliability and validity. Results support the feasibility of using LLMs for scalable creativity assessment in early education contexts. To address this goal, we address two research questions: (1) Criterion-based validity: How well do automated flexibility and originality scores correlate with human ratings? And (2) Content validity: To what extent does the automated model reflect the breadth and diversity of human-defined categories and response distributions in scoring flexibility and originality?

2. Methods

2.1 Population and Data Collection

Participants were 106 third- and fourth-grade students from a diverse urban school in Israel. Of these, 54 were girls, 49 boys, and 3 identified otherwise; 49 were in third grade and 57 in fourth grade. Participants completed the AUT using three objects—chair, pencil, and cup—within 15 minutes, or 20 minutes for those eligible for time accommodations. This study focuses exclusively on responses to the object chair (N = 575 responses), which was the first of the three presented. On average, students provided 5.41 responses (SD = 3.52) to this object. Selecting only the first object helped minimize potential order effects, such as fatigue or learning, that might influence response patterns in subsequent items. Responses were in Hebrew and automatically translated to English. Prior work on flexibility scoring (Hadas & Hershkovitz, 2024) found r = .79 with human translation, suggesting modest noise.

To compare the LLM-based model to human scoring, three human raters, all researchers with expertise in creativity assessment, evaluated flexibility and originality. For flexibility, the raters reached consensus by collaboratively defining a category set and then assigning each response to the most appropriate category. For originality, raters independently scored a random sample of 100 responses using a 1–5 scale (with "1" being "Not Original at All", and "5" being "Very Original"), and the final score per response was calculated as the average across raters. Inter-rater reliability for originality human scores, measured by Fleiss' Kappa, κ = .28 (fair agreement) and pairwise Cohen's Kappa of the three human raters, κ = .18, .25, and .57, lie within the typical span for DT assessments (Silvia, 2011).

2.2 Automatic Flexibility Scoring

For flexibility, we start with an object and its responses. These are used as input to a prompt for the LLM, which returns a set of semantic categories into which the responses can be clustered. Next, for each response, we prompt the LLM again—this time including the object, the set of categories, and the response—asking it to identify the most suitable category for that response. Finally, for each student, we map their responses to categories and count the number of distinct categories they used. This count is used as the student's flexibility score. The prompts used are shown in Table 1. Multiple runs were employed to test for consistency and yielded almost identical results.

2.3 Automatic Originality Scoring

For originality, we take the object and its responses, and split them into batches of 100 to keep the LLM prompt sizes manageable. Each prompt includes the object, the batch of responses, and a description of the target population (i.e., “3rd- and 4th-grade students”). We prompt the LLM to rate the originality of each response on a whole-number scale from 1 to 5, relative to the object and the population. The entire scoring process was run four times, and each response's final originality score was calculated as the average across runs. The full prompt is shown in Table 2.

Table 1. Prompt for Flexibility

For	Prompt
Category Generation	In a recent study, students were given the AUT test for the following object: <object>. The following responses were given: <responses>. Please examine the responses and determine the distinct categories into which you would assign the responses. Output only the category names. I would like the number of categories to be between 15 and 20.
Category Assignment	In a recent study, students completed the Alternative Uses Test (AUT) for the object: <object>. The following responses were provided: <response>. For each response, identify the single most relevant category from the following list: <categories>. Please output only the name of the most suitable category for each response

Table 2. Prompt for Originality

Prompt
You are an expert in childhood creativity and education. Your task is to evaluate responses from 3rd- and 4th-grade students on the Alternative Uses Test (AUT). Children at this age use **simple words, imaginative thinking, and unexpected connections**. Some responses may seem common at first but are highly original for their **developmental level**. Consider the cognitive abilities of young children when scoring originality, rather than applying adult creativity standards. **Scoring Criteria (1-5 Scale)** 1 - **Very Common:** Almost every child would think of this first. 2 - **Somewhat Common:** A slightly creative but still typical use. 3 - **Moderately Creative:** A response that requires some thinking beyond the obvious. 4 - **Very Creative:** An unusual, playful, and unexpected idea for this age. 5 - **Exceptionally Creative:** A highly imaginative response that very few children would think of.

3. Results

3.1 Criterion-Based Validity: Correlation with Human Scores (RQ1)

The automated model demonstrated a strong correlation with human ratings for flexibility (r = .90, p < .01). For originality, the model also yielded a strong correlation (r = .75, p < .01). These results represent an improvement over the Ocsai model, which produced a lower correlation with human ratings on the same dataset (r = .56, p < .01).

3.2 Content-Based Validity: Coverage of Model Results (RQ2)

For flexibility, we compared the category set created by human raters to that generated by the automated model to assess content validity. Human raters identified 19 categories (e.g., Arts and Crafts, Storage, Tent and Shelter), while the model generated 20 categories (e.g., Artistic and Creative Uses, Storage and Organization Uses, Symbolic and Metaphorical Uses). Out of 575 responses, 446 (77.6%) fall into 12 human-defined categories that have direct counterparts in the automated model. For instance, the human category "Arts and Crafts" aligns with the automated category "Artistic and Creative Uses," "Storage" corresponds to "Storage and Organization Uses," and both "Resting" and "Sitting" map to "Sitting and Resting Uses." Even when categories differ between the two sets, they may still be conceptually related, and the total number of categories can remain the same. This demonstrates content validity by showing that the automated model's categories closely align with human-defined ones, capturing the same conceptual coverage in responses.

To further examine the coverage and distribution of both dimensions as measured by the LLM-based model, their descriptive statistics are shown in Table 3. Flexibility scores were moderately variable (*M* = 4.17, *SD* = 1.97), with scores ranging from 1 to 10 and an interquartile range (IQR) of 2. The distribution was slightly right-skewed (skew = 0.67) with moderate kurtosis (1.69), suggesting a mild peak around the median of 4. Originality scores were more concentrated (*M* = 1.83, *SD* = 0.57), with a narrower range (1 to 3.61) and an IQR of 0.79. The distribution was also slightly right-skewed (.68) and relatively flat (kurtosis = 0.42). Notably, originality scores in this child sample did not exceed 4, whereas in adult samples, scores above 4 have been observed (Organisciak et al., 2023). This likely reflects developmental differences in semantic elaboration and conceptual novelty, rather than a limitation of the model.

Table 3. Descriptive Statistics for Automatic Model Results

	Mean	Std	Median	Min	Max
Flexibility	4.17	1.97	4	1	10
Originality	1.83	.57	1.75	1	3.61
	Q1	Q3	IQR	Skew	Kurt
Flexibility	3	5	2	.67	1.69
Originality	1.40	2.19	.79	.68	.42

4. Discussion

4.1 *Automated Scoring Model*

The high correlations between the automated model's flexibility and originality scores and those assigned by human raters confirm similar findings from adult and adolescent populations (Hadas & Hershkovitz, 2024; Organisciak et al., 2023). We believe that averaging multiple runs for originality improves accuracy, by mitigating rounding limitations and capturing finer-grained judgments, and the implications of this approach should be further studied. These results extend the applicability of automated scoring to younger children, emphasizing its promise in educational contexts where scalability is paramount.

The moderate agreement among human raters—while scoring originality—may seem surprising, but this pattern is well-documented in DT research, where even trained evaluators often show inconsistencies in scoring (Silvia, 2011). Such variability highlights a core challenge in assessing DT reliably and emphasizes the importance of automated, consistent scoring schemes (Beaty & Johnson, 2021; Chou et al., 2024).

4.2 Limitations and Future Work

This study has several limitations. First, it relies on automated translation, which may introduce bias, especially in interpreting nuanced language such as metaphors or culturally specific expressions. Future work could explore multilingual LLMs to reduce this dependency. Second,

the sample was drawn from a single school, limiting generalizability. Broader studies across diverse educational and cultural contexts are needed to validate findings. Finally, while this study focused on quantitative validation, future work should incorporate qualitative analyses of scoring discrepancies to improve fairness and developmental sensitivity, while adhering to ethical guidelines that ensure transparent and age-appropriate use.

5. Conclusions

This study demonstrates the potential of LLM-based scoring for assessing DT in children, offering a scalable and consistent alternative—or companion—to traditional human rating. Strong alignment with human scores and semantic categories supports the model's validity.

LLM-based tools can help educators efficiently assess creativity, gain timely insights into cognitive development, and support personalized instruction. Future work should refine prompts for younger learners, expand validation across tasks and contexts, and ensure ethical, transparent use. More broadly, such tools enable scalable assessment, real-time feedback, and longitudinal tracking—laying the groundwork for adaptive educational interventions at scale.

References

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

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2023). The Machines Take Over: A Comparison of Various Supervised Learning Approaches for Automated Scoring of Divergent Thinking Tasks. The Journal of Creative Behavior, 57(1), 17–36. https://doi.org/10.1002/jocb.559

Chou, E., Fossati, D., & Hershkovitz, A. (2024). A Code Distance Approach to Measure Originality in Computer Programming. Proceedings of the 16th International Conference on Computer Supported Education, 541–548. https://doi.org/10.5220/0012632100003693

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

Gubenko, A., & Houssemand, C. (2022). Alternative Object Use in Adults and Children: Embodied Cognitive Bases of Creativity. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.893420

Guilford, J. P., Christensen, P. R., Merrifield, P. R., & Wilson, R. C. (1978). Alternate uses.

Hadas, E., & Hershkovitz, A. (2024). Using Large Language Models to Evaluate Alternative Uses Task Flexibility Score. Thinking Skills and Creativity, 101549. https://doi.org/10.1016/j.tsc.2024.101549

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, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Organisciak, P., Dumas, D., Acar, S., & de Chantal, P.-L. (2024). Open Creativity Scoring [Computer Software]. University of Denver.

Paulus, D. H. (1970). Computer Simulation of Human Ratings of Creativity. Final Report.

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

Runco, M. A., & Acar, S. (2012). Divergent Thinking as an Indicator of Creative Potential. Creativity Research Journal, 24(1), 66–75. https://doi.org/10.1080/10400419.2012.652929

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. Thinking Skills and Creativity, 6(1), 24–30. https://doi.org/10.1016/j.tsc.2010.06.001

Sio, U. N., & Lortie-Forgues, H. (2024). The impact of creativity training on creative performance: A meta-analytic review and critical evaluation of 5 decades of creativity training studies. Psychological Bulletin, 150(5), 554–585. https://doi.org/10.1037/bul0000432