ChatGPT as a Competent Enough Judge in Validating Responses from a Divergent Thinking Task

Hanna Kucwaj (hkucwaj@swps.edu.pl)
Faculty of Psychology in Krakow, SWPS University
Krakow, Poland

Bartłomiej Kroczek (bartek.kroczek@uj.edu.pl)
Center for Cognitive Science, Jagiellonian University in Krakow
Krakow, Poland

Abstract

The validation of responses in divergent thinking tasks is a critical yet understandardized step that should precede creativity scoring. However, inconsistencies related to human judges in this step may compromise the reliability of the results. This study introduces a systematic approach using ChatGPT to validate responses in the Alternate Uses Task (AUT) and compares its performance against six human judges. Analyzing 1245 AUT responses for common objects, we evaluated validity based on precisely defined criteria. Human judges exhibited significant variability, achieving unanimous agreement for only 58% of responses, while ChatGPT demonstrated significant alignment with human assessments, reflecting a capacity to replicate aggregated human judgment. These findings underscore the potential of Large Language Models to enhance objectivity and reproducibility in creativity research by automating response validation. We advocate for integrating AI-driven validation protocols into divergent thinking response evaluation and emphasize transparent reporting of criteria to advance methodological rigor in the field.

Keywords: Altenrate Uses Task; divergent thinking; creativity; Large Language Models; ChatGPT

Introduction

Divergent thinking is widely recognized as a foundation for various creative tasks (e.g., Torrance, 1966; Runco 2010). It is characterized by the ability to generate multiple, unique solutions to open-ended problems, distinguishing it from convergent thinking, which focuses on finding a single correct answer. The Alternate Uses Task (AUT), is a hallmark assessment tool for measuring divergent thinking. This task requires individuals to think of as many uses as possible for a common object, such as a brick or a paper clip, thereby tapping into their creative potential (Runco & Acar, 2012). Classically, participants' responses are assessed using four criteria: fluency (the number of responses generated), flexibility (the diversity of responses), originality (the unconventionality and rarity of responses), and elaboration (the level of detail in a response). Traditionally, responses on the AUT have been scored manually by human judges—a process that is not only time-consuming and effort-intensive, but also potentially unreliable due to the inherent subjective variability in scoring. The challenges of subjectivity and the unreliability of creativity scoring have been widely recognized in the literature (e.g., see Alhashim et al., 2020; Forthmann et al., 2020, 2021; Forthmann & Doebler, 2022; Reiter-Palmon, Forthmann, & Barbot, 2019).

Fortunately, recent advancements in automated methods for rating responses in divergent thinking tasks, using word embedding methods (e.g., GloVe) and large language models (e.g., ChatGPT by OpenAI), have significantly enhanced creativity research (Beaty & Johnson, 2021; Dumas, Organisciak, & Doherty, 2021). Notably, Organisciak et al. (2023) introduced the Open Creativity Scoring with Artificial Intelligence (Ocsai), a free online tool that represents a major step forward in this field. This scoring system is based on a vast dataset of words, allowing it to differentiate between common and uncommon word associations. It represents both the given object (e.g. a brick) and the participant's response as vectors in a semantic space. By measuring the cosine of the angle between these vectors, the system quantifies how closely related the response is to typical word associations. A larger angle indicates a more original response. Ocsai demonstrates a strong correlation with human ratings when scoring the AUT (Organisciak et al., 2023).

Undoubtedly, the automated approach to rating responses in divergent thinking tasks enhances efficiency while also providing a more consistent and objective evaluation of originality. Nevertheless, one more challenge needs to be addressed in open-ended problems: before assessing responses based on the given criteria, it is necessary to exclude responses that are inadequate or invalid (e.g., incomplete, nonsensical, or difficult to understand). For instance, some responses may be overly generic and applicable to any object (e.g., "selling" or "borrowing"), making them uninformative for assessing creative thinking. Others may be implausible or physically impossible, such as suggesting that a brick could be used to build a functional rocket. Such responses, if not eliminated, would not only boost fluency scores and introduce noise into the originality assessment but could also, in the case of automated scoring, lead to abnormally high creativity scores for items that would otherwise be excluded (e.g., due to invalidity). This step might seem trivial at first, but it is, in fact, another task that requires resources, and its outcome may vary depending on the judge's assessment. In the literature, this aspect of data processing—despite being highlighted as important by Guilford himself and his colleagues (Wilson et al., 1960)—is either not reported at all or covered only vaguely, often by simply stating that only correct responses were counted or that incomplete or inappropriate responses were removed (see Reiter-Palmon et al., 2019).

One way to approach this problem is by introducing, alongside the assessment criteria mentioned earlier, an appropriateness or usefulness criterion, defined as whether an idea is feasible and solves the problem (e.g., Diedrich et al.,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

2015). It is worth noting that appropriateness (sometimes referred to as relevance or usefulness**)** is also a key criterion in Amabile's Consensual Assessment Technique (1982), where responses are evaluated not only for their originality but also for their relevance and suitability to the task at hand. Such a criterion, however noteworthy and important from the perspective of some research questions, only partially addresses problematic responses (e.g., unfeasible ones) by "punishing" them with low usefulness score. Nevertheless, the issue of vague (i.e., requiring interpretation), incomplete (e.g., due to time constraints), or highly unspecific responses (those that cannot be associated with a clear function or use) remains a challenge.

Consequently, we decided to propose criteria for systematically removing such responses and implemented these criteria as a set of instructions for a Large Language Model (LLM) assessment, specifically using ChatGPT by OpenAI. LLMs provide powerful capabilities for analyzing research data in natural language, enabling more efficient and scalable assessments across various domains. It has been shown that LLMs can perform at least as well as humans in several fields, such as providing feedback on writing (Liang et al., 2024), multilingual text analysis (Rathje et al., 2024), and generating responses to patients questions (Ayers et al., 2023). In some aspects, LLMs may provide leverage beyond humans due to their greater reproducibility in structured workflows (Staudinger et al., 2024) alongside reduced susceptibility to context-dependent human biases (Chen et al., 2024). Therefor, we asked the ChatGPT model and six human judges to evaluate whether a response from the AUT is valid (and should be included in future creativity assessment) or whether a response is invalid (i.e., is incomplete, nonsensical, or difficult to understand), based on the same criteria. As a result, we were able to assess the extent to which (1) human judges agree with one another and (2) ChatGPT agrees with the human judges.

Method

Participants

For the purposes of this study, we defined two groups of participants: task participants, who generated alternative uses for three given objects, and human judges, who validated these responses based on given criteria.

Task participants. Fifty participants (15 men, 31 women, 2 preferred not to disclose their gender, and 2 with missing demographic data due to technical issues) were recruited through a university's online recruitment system. Participants' ages ranged from 19 to 51 years (M = 27.19, SD = 8.97; 2 participants had missing age data due to technical issues). All individuals completed the AUT online and received course credit for their participation. All participants provided informed consent to participate and were informed that they could stop the experiment at any time. Data were anonymized. All other aspects of the procedure conformed to the WMA's Declaration of Helsinki.

Judges. Six people (3 female, 3 male) aged between 19 and 37 (M = 25, SD = 6.08), were asked to validate participants' responses from the AUT based on given criteria. They were recruited from a university student population and received course credit for completing the task.

Alternate Uses Task

The task consisted of three items for which participants generated alternate uses: a paperclip, a rubber band, and a newspaper. The instruction was as follow: "This task requires creative thinking. Your job will be to come up with as many and as diverse uses as possible for a given object. You will have 3 minutes for each object. Both the number of ideas and their originality are important. Try to think creatively and unconventionally". Participants inserted their response in a text box, and were asked to separate each idea with a colon. As a result, we obtained three files—one for each object—containing two columns: participant IDs and the alternative uses provided by them, without any corrections or modifications, in their native language (all participants shared the same nationality). Each file served as input data for both ChatGPT and human judges.

Instruction for ChatGPT

In this paragraph, we will describe each command used to build a prompt comprising an instruction for ChatGPT to perform validation of responses from the AUT. The entire prompt content is available online at https://osf.io/jazem.

Step 1: Semicolon separation. As the AUT is open-ended, ill-structured task, participants differ substantially in their response format (some people did not strictly follow the instruction to separate ideas with a colon; for instance, they numbered their ideas instead). First, we prompt ChatGPT to unify the format of responses by removing numbers, excessive punctuation marks, and unnecessary spaces between expressions. Consistent with prompt engineering principles aimed at enhancing model reliability (e.g., specifying output structure, minimizing ambiguity, and emphasizing task-specific constraints), the instructions prioritized simplicity and iterative refinement to ensure that ChatGPT's processing aligned with the study's objectives. As a result, we obtained cleaned lists of ideas separated by semicolons. We verified ChatGPT's accuracy in this step and found that in only 1.36% of cases, ChatGPT made errors by splitting one response into multiple responses or by combining two responses into one, which we deemed sufficiently accurate to proceed.

Step 2: Translating responses into English. To fully utilize the capabilities of ChatGPT, we decided to translate the original responses from Polish into English. This decision aligns with research showing that LLMs exhibit significantly better performance in high-resource languages like English due to their disproportionate representation in training corpora (Huang et al., 2023; Li et al., 2024). Importantly, Zielińska et al. (2023) demonstrated that translating responses from Polish into English did not substantially affect the scoring accuracy of the Ocsai model. As a result, we obtained the lists from Step 1 but translated into English. The translation was obtained using Google Translate. For the same reason, only the command for Step 1 was written in the language consistent with the participants' responses. The rest of the commands for ChatGPT were written in English.

Step 3: Validation of responses. This step is essential, as it involves determining which responses should be classified as "meeting criteria" and which should be removed from the dataset as "not meeting criteria" before further creativity scoring (which was not a topic of this study). The task was to: (1) validate each idea from the lists created in Step 2 based on the given criteria (see Table 1), and (2) for each item flagged as "does not meet criteria" provide feedback explaining the decision (see Table 2 for feedback examples provided to ChatGPT; for each object, between 7 and 9 examples were supplied). Requiring a rationale for each decision follows prompt engineering recommendations, where asking models to explain their reasoning improves accuracy, transparency, and consistency. This approach aligns with best practices for working with LLMs, specifically Chain-of-Thought (CoT) prompting, as provided by OpenAI (2022) and Wei et al. (2022).

Apart from the validation criteria and example evaluations, at the end of the prompt we included a reminder for ChatGPT to be concise and clear, to allow for unethical uses (since ChatGPT is typically reluctant to provide information that could potentially cause harm or pose ethical concerns), to focus on the criteria, and to avoid providing additional content that has not been requested.

As a result, we obtained two lists for each object: one list with uses validated as meeting the criteria and second list with uses validated as not meeting criteria with feedback for each excluded use.

Step 4. Decreasing validation variability. LLMs are primarily constructed to provide various responses and generate diverse, contextually relevant outputs. Consequently, even when our validation criteria were explained in great detail, some particularly difficult-to-classify responses were occasionally classified differently across separate runs. To reduce variability in ChatGPT's outputs, we took two actions: (1) we adjusted the model's parameters (see the next paragraphs), and (2) we ran Step 3 three times and prompted ChatGPT to perform a voting process. This meant that in edge cases (i.e., difficult-to-classify), the response that received the majority classification was selected as the final output. As a result, we obtained the same output as at the end of Step 3 (two lists for each object: "meets criteria" and "does not meet criteria"), but with minimized variability.

Table 1: Validation criteria provided in both ChatGPT prompt and the instruction for human judges. The criteria were the same for each object, i.e., a rubber band, a paperclip, a newspaper. Instead of the word *object*, the name of the specific item was inserted.

Criterion	Explanation
Direct Relevance to the Object	The suggested use must directly involve the *object* and utilize its inherent properties such as elasticity, size, material, or functionality.
Physical Plausibility	The use should be physically possible without violating the laws of physics or requiring supernatural elements. Allow for uses that are potentially unsafe, inappropriate, unethical, or even dangerous, as long as they are theoretically possible.
Specificity and Detail	Responses should be clear and specific. Vague or overly broad statements are unacceptable.
Minimal Modification Requirement	The suggested use should be feasible with the *object* in its standard form or with minimal alterations that do not fundamentally change its nature.
Functional Feasibility	The use should make logical sense and serve a practical purpose, even if unconventional.

ChatGPT model setup

The GPT-4o-2024-08-06 model was configured with specific parameters to enhance replicability and consistency. A seed value of 12345 was applied to maximize the determinism of outputs and facilitate reproducibility across runs. To minimize randomness and promote focused responses, the top-p value was set to 0.1, and the temperature was adjusted to 0.1. While newer model versions may offer enhanced capabilities, their adoption should not be taken for granted. It is crucial to thoroughly evaluate them to ensure they align with task-specific requirements, as they may paradoxically perform worse than previous models, potentially compromising reliability.

Instruction for human judges

The judges were asked to follow the instructions provided in Polish. Crucially, the content of instruction was exactly the same as the instructions given to ChatGPT, with only one exception as human judges did not proceed with step 2 since they performed their assessments on the original, non-translated responses, and step 4. The form of a final outcome of their assessment was the same as in the case of ChatGPT, which are the files for each object with two lists: one list with uses validated as meeting the criteria and second list with uses validated as not meeting criteria with feedback for each excluded use. Each judge was instructed to validate files containing responses for a paperclip, a rubber band, and a newspaper in a different order.

Table 2: Example evaluation for a rubber band provided in both ChatGPT prompt and the instruction for human judges. For each object between 7 and 9 examples were provided.

Alternative Use	Evaluation	Feedback
Hair tie	Meets criteria	Using a rubber band to tie back hair utilizes its elasticity and holds hair securely.
Slingshot	Meets criteria	Creating a simple slingshot with a rubber band employs its stretchiness to launch small objects.
Weapon	Does not meet criteria	A rubber band can be used as a projectile launcher, such as in a homemade slingshot. However, simply a 'weapon' doesn't capture its specific functionality and is too vague.
Replacing a broken fan belt	Does not meet criteria	This use fails the Physical Plausibility and Minimal Modification Requirement criteria. A rubber band cannot withstand the mechanical stress of a fan belt and would not function safely or effectively.

Results

Overall, participants provided 1245 responses in the AUT for judges' evaluation. On average, our judges evaluated 83% of these responses as meeting the given criteria, suggesting that, in general, participants generated valid responses in the AUT. However, all six judges agreed on the same response status in only 58% of cases—either all validated a use as 'does not meet criteria' (first bar in Figure 1) or as 'meets criteria' (last bar in Figure 1). This highlights concerns about reproducibility related to subjective judgment and the need to address this issue. Figure 1 illustrates the distribution of responses validated as "meets criteria" based on the number of judges who classified them as such.

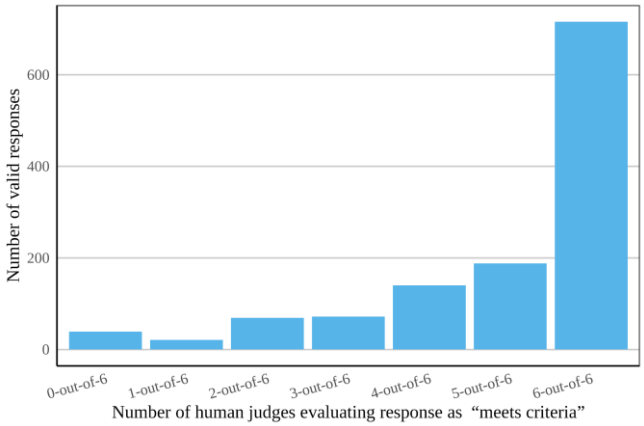

Figure 1: Bar chart illustrating the number of valid responses as evaluated by human judges. The x-axis represents the number of human judges (out of six) who rated a response as meeting criteria. The y-axis shows the corresponding number of valid responses.

Consistency among the judges

Human judges made binary decisions ("meets criteria" vs. "does not meet criteria") for all participants' responses (uses) in the AUT. First, we conducted Cochran's Q test to assess whether significant differences existed in the patterns of exclusion decisions for the same use across judges. A significant result ($p < .05$) would indicate systematic inconsistencies between judges' assessment. Following a significant Q test, pairwise McNemar tests with Holm-Bonferroni correction were applied to all judge pairs to identify specific disagreements. McNemar's test evaluated marginal homogeneity in 2×2 contingency tables. Cochran's Q test revealed statistically significant differences in validation rates across judges, $Q(5) = 156.48$, $p < .001$. This indicates that the likelihood of an option being validated as meeting criteria systematically differed among the judges, rather than occurring by chance. Post hoc pairwise comparisons using McNemar's tests with Holm-adjusted p-values further supported this finding: of the 15 possible pairs-of-judges comparisons, a majority (13 comparisons, 86.7%) yielded statistically significant differences at the $\alpha = .05$ level. These results collectively suggest substantial heterogeneity in validation tendencies across judges, with systematic disagreements extending beyond random variation. A smaller subset of comparisons (2 comparisons, 13.3%) did not reach statistical significance after adjustment, indicating no evidence of disagreement between specific pairs of judges. Overall, the pattern strongly supports the conclusion that validation outcomes were inconsistently applied across the group of our judges.

Consistency between the human judges and ChatGPT

To evaluate whether ChatGPT validates responses similarly to human judges, we employed the Jonckheere-Terpstra test, Kendall's rank correlation, and linear regression in our analyses. The Jonckheere-Terpstra test was used to formally test whether the relationship between the ordinal variable (judges' consistency) and the continuous variable (proportion of ChatGPT responses evaluated as 'meets criteria') was

monotonically increasing. Kendall's rank correlation provided a measure of the strength of this effect. Finally, linear regression was included as a widely recognized tool allowing for further exploration of the relationship between the variables.

The Jonckheere-Terpstra test revealed a significant ordered trend in the data, JT = 323002, p < .001, indicating a systematic relationship between the number of human judges evaluating response as "meets criteria" and ChatGPT's evaluation of the same responses. This finding was further supported by Kendall's rank correlation analysis, which demonstrated a moderate positive association between human validation and ChatGPT's validation, τ = .436, z = 16.61, p < .001. This suggests that ChatGPT's validation patterns align with those of human judges to a meaningful degree (see Figure 2).

A linear regression was conducted to quantify the linear relationship between human judges' consistency and ChatGPT's validation. Since human judges' consistency is an ordinal factor, it was converted into a numeric variable for the regression analysis. In this representation, 0 indicates that none of the human judges evaluated a response as "meets criteria" (i.e., all judges unanimously decided to exclude the response), 2/6 indicates that two out of six judges evaluated the response as "meets criteria" while four evaluated it as "does not meet criteria", and 1 (six out of six) indicates that all judges agreed that the response met the criteria for inclusion. The model was statistically significant, F(1, 1243) = 339.40, p < .001, with human judges consistency explaining 21.5% of the variance in ChatGPT's validation decisions (R^2 = .215, adjusted R^2 = .214). Both the intercept (b = 0.142, SE = 0.035, p < .001, 95% CI [0.072, 0.211]) and the slope (b = 0.752, SE = 0.041, p < .001, 95% CI [0.672, 0.832]) were statistically significant (see the regression line at Figure 2). The residual standard error was 0.377, suggesting moderate variability in ChatGPT's decisions not explained by the model.

Overall, these results indicate that ChatGPT's validation aligns significantly with that of human judges, with the proportion of human judges rating a response as 'meets criteria' serving as a meaningful predictor of ChatGPT's validation patterns. This suggests that ChatGPT demonstrates a validation process that is broadly consistent with human judgment.

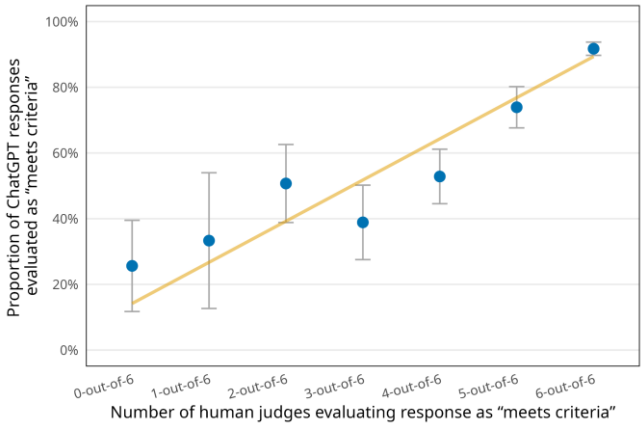

Figure 2: Percentage of participants' responses validated as "meets criteria" by ChatGPT, based on the number of human judges who evaluated the responses as meeting the criteria. The x-axis represents the number of human judges (out of six) who rated the responses as "meets criteria," while the y-axis shows the percentage of participants' responses validated by ChatGPT as meeting the criteria. Blue dots indicate the observed percentages, with error bars representing 95% confidence intervals. The orange line represents a linear trend, illustrating the positive relationship between the number of human judges who validated a response as "meets criteria" and the percentage of responses validated by ChatGPT as such.

Discussion

The aim of the present study was to propose a systematic method to validate responses in the AUT—an open-ended, divergent thinking task—before proceeding with creativity scoring. Unlike creativity scoring itself, this step has rarely been a topic of research concern, even though it may influence the results of divergent thinking tasks. Importantly, our approach is based on an AI method, which fits into the broader context of changes in the field of creativity scoring and measurement. To validate our method, we asked six human judges and ChatGPT to decide whether participants' responses in the AUT meet or do not meet given criteria. First, we assessed the extent to which human judges were consistent in their evaluations. Second, we investigated whether ChatGPT could serve as a judge in performing this task in future. In the following section, we will present these findings in detail.

First, the study revealed significant variability among human judges in evaluating responses, with only 58% of cases achieving unanimous agreement regarding classification as "meets criteria." Pairwise comparisons between judges demonstrated high rates of disagreement, with most judge pairs diverging in their classifications. These findings suggest that subjective interpretation and individual biases substantially influenced validation outcomes, raising concerns about the reproducibility of results. Second, we found that ChatGPT's validation patterns were significantly aligned with those of human judges, reflecting a meaningful level of agreement between the two.

Consequently, we suggest that ChatGPT is a competent enough judge in validating responses from a divergent thinking task. Moreover, we argue that delegating this task to

ChatGPT presents distinct advantages for creativity research. This approach demonstrates potential to enhance reproducibility and transparency through standardized implementation protocols, while substantially reducing the temporal, financial, and labor resources typically required for human-driven processes.

In summary, we propose integrating LLMs into this often-overlooked step of analyzing responses in divergent thinking task, offering a promising solution to streamline the validation process.

Future directions

Importantly, we do not claim that the list of criteria or examples we formulated for ChatGPT and human judges is fixed or absolute. On the contrary, we expect these criteria may vary depending on the authors' assumptions and the specific goals of the task. What we emphasize, however, is the importance of systematically reporting this validation stage before proceeding with creativity scoring. This step of validating and rejecting incomplete, invalid, or unclear responses is crucial, and we encourage a transparent and detailed account of how it is handled.

Moreover, current advancements in LLMs suggest that ongoing improvements may necessitate a paradigm shift in addressing this issue. One emerging trend supporting this possibility involves the increased adoption of smaller, highly specialized language models trained explicitly for narrow task domains. Such specialization could be achieved through either de novo system design or, more plausibly, via fine-tuning existing general-purpose LLMs (e.g., ChatGPT) on domain-specific datasets to enhance task-specific performance. For example, similar approach has been adopted in developing Ocsai (Organisciak et al., 2023).

Limitations

The development of AI models such as ChatGPT represents a challenge in their application to problem-specific prompts. While these models systematically improve, their rapid evolution simultaneously necessitates ongoing adjustments to prompting strategies, as methods optimized for earlier iterations risk obsolescence in newer versions. Notably, older models may occasionally outperform their successors for niche applications, underscoring the need for flexibility in model selection (Nikolic et al. 2023).

A critical limitation stems from the inherent "black-box" nature of AI systems. Their opaque internal mechanisms preclude direct insight into evaluating processes, rendering prompt engineering an iterative, trial-and-error endeavor. Even with established guidelines for prompt design, achieving optimal outputs often requires repeated refinement.

Although ChatGPT's validation decisions could theoretically reflect biases from its training data, its vast dataset may offer a statistical normalization—potentially mitigating certain outlier or individual-specific biases more broadly than a single human judge or a group of judges.

Taken together, these considerations highlight both the practical challenges and the potential robustness of AI-assisted evaluation. Our approach provides valuable insights into the role of AI in creativity research, and ongoing advancements in AI models and prompt engineering could further enhance the reliability and adaptability of this method.

Conclusions

Our study demonstrates a systematic AI-based method for validating responses in the AUT before proceeding with creativity scoring. The analysis of human evaluations revealed significant variability, highlighting the need for standardized validation processes. Based on these insights, we suggest that ChatGPT could provide a solution, as it not only demonstrates strong alignment with human evaluators but also offers distinct advantages. While we do not claim this is the only viable approach to response validation, our goal was to offer an illustrative example of how large language models can be integrated into creativity research to support greater objectivity, consistency, and scalability.

Acknowledgements

This work was supported by the National Science Center, Poland [Grant Number: 2021/41/N/HS6/01620].

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Nobles, A. L., & Longhurst, C. A. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596. doi:10.1001/jamainternmed.2023.1838

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43, 997-1013. doi:10.1037/0022-3514.43.5.997

Alhashim, A., Marshall, M., Hartog, T., Jonczyk, R., Dickson, D., Van Hell, J., Okudan-Kremer, G., & Siddique, Z. (2020). Work in Progress: Assessing Creativity of Alternative Uses Task Responses: A Detailed Procedure. *2020 ASEE Virtual Annual Conference Content Access Proceedings*, 35612. doi:10.18260/1-2--35612

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. doi:10.3758/s13428-020-01453-w

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic Distance and the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Creativity Research Journal*, 34(3), 245–260. doi:10.1080/10400419.2022.2025720

Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or LLMs as the judge? A study on judgement biases. arXiv (Preprint). http://arxiv.org/abs/2402.10669

Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., & Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12365–12394). doi:10.18637/jss.v067.i01

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*(1), 35–40. doi:10.1037/a0038688

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. doi:10.1037/aca0000319

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. doi:10.1037/aca0000518

Forthmann, B., Jankowska, D. M., & Karwowski, M. (2021). How reliable and valid are frequency based originality scores? Evidence from a sample of children and adolescents. *Thinking Skills and Creativity*, 41, 100851. doi:10.1016/j.tsc.2021.100851

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*, 90(3), 683–699. doi:10.1111/bjep.12325

Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., & Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12365–12394. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-emnlp.826

Johnson, D.R., Kaufman, J.C., Baker, B.S. *et al.* Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods* 55, 3726–3759 (2023). doi:10.3758/s13428-022-01986-2

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., & Du, M. (2024). Language Ranker: A metric for quantifying LLM performance across high and low-resource languages. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2404.11553

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., ... & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), (Preprint) arXiv:2310.01783

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., ... & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multistakeholder study. *Australasian Journal of Engineering Education*. doi:10.1080/22054952.2024.2372154

OpenAI. (2022). *Best practices for prompt engineering with OpenAI models.* Retrieved from https://platform.openai.com/docs/guides/prompt-engineering

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. doi:/10.1016/j.tsc.2023.101356

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34), e2308950121. doi:10.1073/pnas.2308950121

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. doi:10.1037/aca0000227

Runco, M. A. (2010). Divergent thinking, creativity, and ideation. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (pp. 413–446). Cambridge University Press. doi:10.1017/CBO9780511763205.026

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, 24(1), 66–75. doi:10.1080/10400419.2012.652929

Staudinger, M., Kusa, W., Piroi, F., Lipani, A., & Hanbury, A. (2024). A reproducibility and generalizability study of large language models for query generation. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 186–196. doi:10.1145/3673791.3698432

Torrance, E. P. (1966). *Torrance test of creative thinking: Norms-technical manual research edition-verbal Tests, forms A and B-figural tests, forms A and B.* Princeton: Personnel Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. (Preprint). arXiv:2201.11903

Wilson, R. C., Christensen, P. R., Merrifield, P. R., & Guilford, J. P. (1960). Alternate Uses–Form A: Manual of administration, scoring, and interpretation (2nd preliminary ed.). Beverly Hills, CA: Sheridan Supply Company.

Zielińska, A., Organisciak, P., Dumas, D., & Karwowski, M. (2023). Lost in translation? Not for Large Language Models: Automated divergent thinking scoring performance translates to non-English contexts. *Thinking Skills and Creativity*, *50*(101414), 101414. doi:10.1016/j.tsc.2023.101414