ISSN 2616-7433 Vol. 8, Issue 4: 28-32, DOI: 10.25236/FSST.2026.080405

Designing Divergent-Thinking Tasks for AI-Based Scoring Algorithms

Bin Wang

Department of Computer Science, School of Intelligent Science and Engineering, Xi'an Peihua University, Xi'an, Shaanxi, China
bin.wang@peihua.edu.cn

Abstract: *Automated creativity scoring is spreading through schools as advice about choosing a tool, yet any tool, however well validated, can only score fairly what a task elicits. This article addresses the other half of the practice: the design of divergent-thinking tasks for AI-based scoring algorithms. Six task-design recommendations are proposed and developed: anchoring the task in an established divergent-thinking family, matching the response format to the declared capability of the scoring tool, fixing a level-anchored rubric before responses are collected, stating the unit of scoring and bounding response length, piloting the task against a small human-scored sample, and using fresh, curriculum-local item content. Each recommendation names the action, the rationale, and the consequence for score interpretation. A worked example carries one task from design brief to classroom deployment, and a closing section marks the limits of task-side effort. A task built this way gives a validated tool something meaningful to score and gives teachers a stable basis for reading the numbers.*

Keywords: *divergent thinking; creativity assessment; task design; automated scoring; artificial intelligence in education*

1. Introduction

Schools that adopt automated creativity scoring concentrate their scrutiny on the choice of tool: which system, which vendor claims, which price. That scrutiny is earned, because the differences between scoring systems are real and the evidence behind them varies widely. Yet a second decision shapes the outcome just as much and receives far less attention: the design of the task whose responses will be scored. An artificial intelligence system that scores open-ended responses does not create meaning. It attaches numbers to whatever the task produced. Given a vague prompt, such a system will score noise with impressive speed. Given a task with a construct behind it, a scale fixed in advance, and responses in a format it can read, the same system produces evidence a teacher can act on.

Task design is also the decision that stays with the school. Tools are bought once and switched rarely; tasks are written lesson by lesson, usually by the teacher, and they change with every unit of instruction. A school that designs tasks well gets more from a modest tool than a school that designs tasks badly and buys the strongest system on the market. Task design is, in this sense, the cheaper and more durable half of the practice.

This article sets out six task-design recommendations for divergent-thinking tasks intended for automated scoring, aimed at educators without measurement training. The recommendations do not assume any particular tool and apply to any system that scores text responses against a scale. The article develops each recommendation with its rationale and its characteristic failure mode, carries one complete task through all six recommendations as a worked example, and closes with the limits of what task-side effort can achieve.

2. Grounding Score Meaning in the Task

Divergent-thinking tasks ask for many varied ideas from a single starting point, and they carry the longest scoring tradition in creativity assessment. Guilford's divergent-production factors, fluency, flexibility, originality, and elaboration, defined what such tasks elicit and gave the resulting scores their meaning [1]. The Torrance tests packaged the paradigm into school-usable forms with published scoring manuals, and five decades of application have made them the reference point of classroom

ISSN 2616-7433 Vol. 8, Issue 4: 28-32, DOI: 10.25236/FSST.2026.080405

creativity measurement [2]. Divergent thinking remains the most practical available indicator of creative potential: broad enough to train and measure, specific enough to score [3].

The tradition is not merely old; it carries evidence of consequence. A fifty-year follow-up of Torrance-test participants traced childhood scores to adult creative achievement in occupations, publications, and public works, which is the kind of long-range validity no automated system can yet claim [4]. Score trends on these instruments have even entered policy discussion, with analyses of normative data reporting declining scores across successive cohorts [5]. A school that designs its own tasks inherits this evidence base when it stays inside the established families, and forfeits it when it does not.

The score of a divergent-thinking response therefore does not belong to the response alone. It results from a task, a scale, and a scoring procedure agreed in advance. Human scoring of creative products holds together only under disciplined conditions: the Consensual Assessment Technique reaches agreement when qualified judges rate products independently against a shared sense of the domain [6], and studies of divergent-thinking scoring find that brief instructions bring even novice raters to adequate agreement, which is what makes classroom scoring feasible at all [7]. Change the procedure and conclusions about the same responses can change; the originality of identical answer sets reads differently under different scoring techniques [8]. These dependencies are not weaknesses of the tradition. They are the mechanism by which the tradition means anything.

Automated scoring inherits every one of these dependencies and adds a format constraint. The automation of idea scoring has a longer history than the current wave of language-model systems suggests: latent semantic analysis introduced the measurement of semantic distance between an idea and its prompt as a computable quantity a quarter of a century ago [9], and open research platforms such as SemDis made semantic-distance scoring available to laboratories that never trained a rater [10]. Large language model systems now score open-ended text against an instructor-supplied rubric at classroom scale, which moves the scoring procedure into the prompt and the task materials. Whatever scores the response, the task still decides what the number can mean. A well-known item invites rehearsed answers. A rubric written after the fact bends to the responses it was meant to judge. An unbounded response format mixes one-line and page-long answers into one scale. Fairness on the scoring side and fairness on the task side are two halves of one claim, and only the second half is under the school's direct control.

3. Developing the Task-Design Recommendations

Table 1 states the six recommendations. The subsections develop each group in turn: anchoring the construct, matching task and tool, fixing the scale and the unit, and the two recommendations the school can verify internally before deployment.

Table 1. Six task-design recommendations for divergent-thinking tasks intended for automated scoring.

Recommendation	What to do	Why it matters
T1 Anchor the construct	Use an established divergent-thinking family (uses for objects, consequences, product improvement) with its standard instructions	The task family carries decades of scoring practice; free-form prompts leave the score without a construct behind it
T2 Match format and tool	Collect responses the tool can read: typed text in one language for text-only tools; pictures or handwriting only if the tool declares them	A score produced from material the tool cannot parse, or was never validated on, is not evidence
T3 Fix the scale first	Write a level-anchored rubric, with named levels, before collecting responses, and supply it to the scoring tool	A scale written after seeing the responses bends to them; anchors make scores comparable across classes and terms
T4 State the unit and bounds	Decide per-idea or whole-response scoring, and set a response length range	Idea counts reward length; whole-answer scoring rewards elaboration; mixed units make scores mean different things
T5 Pilot with human scores	Score 20 to 50 responses by hand, then have the tool score the same set, and compare before deployment	The comparison is the only direct evidence the tool reads this task the way a trained reader does
T6 Use fresh item content	Build items from curriculum-local objects and current material rather than famous test items	Well-known items circulate with their sample answers; new content makes the tool meet student thinking

ISSN 2616-7433 Vol. 8, Issue 4: 28-32, DOI: 10.25236/FSST.2026.080405

3.1 Anchoring the Construct in Established Families

The established families earned their place through use. Uses-for-objects, consequences, and product-improvement prompts arrive with instructions that generations of studies have employed, so the meaning of the resulting scores is anchored in shared practice rather than in the intentions of whoever wrote this week's worksheet [1-3]. A teacher adapting the families keeps their structure and varies only the content: unusual uses for an object from the current unit of study is a sound item; invent a new prompt format mid-semester is a research project.

The characteristic failure of this recommendation is prompt drift. A task that begins as uses-for-objects acquires extra constraints over iterations (must be physically possible, must not repeat materials, must be funny), each of which changes what fluency, flexibility, and originality measure. When adaptation is needed, the standard instructions should be reproduced verbatim and the variation confined to the object named, which keeps the score comparable to the tradition the school is borrowing.

3.2 Matching Task Format and Tool Capability

Every scoring system reads certain formats and ignores or misreads the rest. Text-only tools require typed responses in the language the tool supports; figural or handwritten responses need a tool that declares them. The PISA 2022 creative-thinking assessment illustrates the discipline at policy scale: its framework separates written and visual response formats precisely so that each is scored by a procedure suited to it [11]. The design recommendation is a single question answered before the lesson is announced: will every student response arrive in a format this tool was validated on? Where the answer is no, the task changes, not the expectation. Asking a text-only tool to score photographs, or a tool validated on English to score Chinese, converts the instrument into a random number generator wearing the same interface.

3.3 Fixing the Scale and the Scoring Unit

A rubric written before data collection is a measurement instrument; a rubric written after is a description of the data with ambitions. Named levels, from commonplace to exceptional with the criteria for each, give the automated tool a scale with interpretable steps and give the teacher stable categories for reading results across classes and terms. The unit of scoring belongs in the same decision. Scoring each idea separately rewards fluency, since more ideas mean more chances at a high rating; scoring the response as a whole rewards elaboration and integration. Either is defensible when stated in advance, and neither can be inferred later from the same numbers. A response-length range completes the fixation. Without one, a three-page essay and a one-line quip share a scale, and the tool has no stable basis for preferring either.

3.4 Piloting the Task and Refreshing the Content

The last two recommendations need no external expert, no vendor cooperation, and no statistical training beyond a scatter of agreement. For the pilot, the teacher scores a small sample of responses by hand with the rubric, then has the tool score the same sample. Reading the disagreements is the work: a tool that systematically underrates the teacher's top band, or that clumps every response into two adjacent levels, has told the school something it needs to know before the scores reach a gradebook. The sample of 20 to 50 responses is large enough to reveal a systematic gap and small enough to fit one free period.

Content freshness is the quieter recommendation. Widely reproduced test items travel with their sample answers, and a language-model-based scorer may have met those answers during training. An item built from an object in this term's curriculum, or from a material specific to the school's own context, returns the task to the students' thinking, which is what the school wanted to measure in the first place. Freshness also disciplines the reuse of last year's items, since a task that circulates among cohorts gradually measures recall of the marking scheme rather than ideation.

4. Working through One Task: From Design Brief to Deployment

Consider a design-and-technology teacher who wants an AI-scored creativity task for a Grade 8 unit

ISSN 2616-7433 Vol. 8, Issue 4: 28-32, DOI: 10.25236/FSST.2026.080405

on materials. The brief: one class period, scores reported as formative feedback within the week.

The teacher anchors the task in the uses-for-objects family and names an object from the unit, an empty marker pen (T1). The instructions are the standard ones for the family, reproduced verbatim from the scoring tradition: list as many unusual uses as you can; quantity and variety count; do not worry about practicality. The response format is fixed next (T2): students type their answers into the learning platform in Chinese, one use per line, because the school's scoring tool is a text-based system whose documentation reports validation on Chinese responses. A drawn alternative is offered to no one; the drawing version of the task is a different task.

The rubric is written before the lesson and supplied to the tool with the task (T3). Four named levels range from conventional (uses that keep the object's function or restates it) through reinterpretation (the object serves a different system) to exception (a use that changes the frame of the problem itself). The unit of scoring is the individual idea, and the response bound is three to ten ideas of roughly one line each (T4), so that fluency is scored as intended and a long essay cannot masquerade as many ideas.

After the lesson, the teacher hand-scores thirty responses with the rubric before the tool sees them, then runs the same thirty through the tool and compares (T5). The two score sets track closely on the middle levels, while the tool rates two of the teacher's top-band responses one level lower; the teacher reads both responses, finds the tool's reading defensible for one and narrow for the other, and records the pattern as a known lean of this tool on frame-changing ideas. The item content itself was drawn from the current unit and has not been used in published materials the students could have seen (T6).

Deployment follows with eyes open. Scores are reported as a band with a comment, not a bare number, and the recorded lean of the tool is mentioned to the class as part of learning how scores come to be. The whole cycle cost the teacher one preparation period and half a free period, and it produced, as a by-product, the only local validity evidence the school owns: a documented comparison between the tool's reading and a trained human reading of its own task.

5. Bounding the Claims

Task design cannot rescue an unvalidated tool, and tool validation cannot rescue a badly designed task; the six recommendations in this article and the tool-side evidence they presuppose are complements, not substitutes. Three further cautions bound the claims. First, automated scores remain estimates: agreement between the tool and a human reader, even where demonstrated, carries a band of uncertainty that a bare integer conceals, and reporting should preserve the band rather than the point. Second, repeated use of the same fresh item erodes its freshness; a task bank with rotation, and items retired after a term, keeps the measurement ahead of student memory. Third, tasks administered across language backgrounds mix measurement with proficiency: a student scoring low on fluency in a second language may be displaying vocabulary access rather than ideation, and the task-side remedies, such as allowing responses in the stronger language when the tool permits, are design decisions of exactly the kind this article is about. None of these cautions weakens the central claim. They mark the territory within which the six recommendations deliver what they promise: interpretable, comparable, and fair-to-a-defensible-degree scores from a tool the school has reason to trust.

6. Conclusion

Choosing a scoring tool and designing the tasks are the two decisions a school controls, and the second deserves the same discipline as the first. The six recommendations hold the construct anchored, the format matched, the scale and unit fixed before responses exist, and the tool verified on the school's own task with the school's own students. A task built this way does not make an unvalidated tool trustworthy. It makes a validated tool's scores interpretable, comparable across classes and terms, and worth the teacher's attention, and it leaves the school owning local evidence no vendor demonstration can supply. Task design is where a school's professional judgment enters automated creativity assessment, and it is the half of the practice that cannot be outsourced.

Acknowledgments

This work was supported by the 2024 Shaanxi Provincial Education Science Planning Project for the 14th Five-Year Plan, grant number SGH24Y2697.

ISSN 2616-7433 Vol. 8, Issue 4: 28-32, DOI: 10.25236/FSST.2026.080405

References

[1] Guilford, J. P. (1967). The Nature of Human Intelligence. McGraw-Hill.
[2] Torrance, E. P. (1974). Torrance Tests of Creative Thinking: Norms-Technical Manual. Scholastic Testing Service.
[3] Runco, M. A., & Acar, S. (2012). Divergent Thinking as an Indicator of Creative Potential. Creativity Research Journal, 24(1), 66-75.
[4] Runco, M. A., Millar, G., Acar, S., & Cramond, B. (2010). Torrance Tests of Creative Thinking as Predictors of Personal and Public Achievement: A Fifty-Year Follow-Up. Creativity Research Journal, 22(4), 361-368.
[5] Kim, K. H. (2011). The Creativity Crisis: The Decrease in Creative Thinking Scores on the Torrance Tests of Creative Thinking. Creativity Research Journal, 23(4), 285-295.
[6] Amabile, T. M. (1982). Social Psychology of Creativity: A Consensual Assessment Technique. Journal of Personality and Social Psychology, 43(5), 997-1013.
[7] 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.
[8] Plucker, J. A., Qian, M., & Wang, W. (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.
[9] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. Discourse Processes, 25(2-3), 259-284.
[10] 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.
[11] OECD. (2023). PISA 2022 Assessment and Analytical Framework. OECD Publishing.