



Towards Game-Based Assessment of Creative Thinking

Janet Rafner, Qian Janice Wang, Miroslav Gadjacz, Thomas Badts, Brendan Baker, Carsten Bergenholtz, Michael Mose Biskjaer, Thomas Bui, Andrea Carugati, Matthieu de Cibeins, Lior Noy, Seyedahmad Rahimi, Kristian Tylén, Blanka Zana, Roger E. Beaty & Jacob Sherson

To cite this article: Janet Rafner, Qian Janice Wang, Miroslav Gadjacz, Thomas Badts, Brendan Baker, Carsten Bergenholtz, Michael Mose Biskjaer, Thomas Bui, Andrea Carugati, Matthieu de Cibeins, Lior Noy, Seyedahmad Rahimi, Kristian Tylén, Blanka Zana, Roger E. Beaty & Jacob Sherson (2023) Towards Game-Based Assessment of Creative Thinking, Creativity Research Journal, 35:4, 763-782, DOI: [10.1080/10400419.2023.2198845](https://doi.org/10.1080/10400419.2023.2198845)

To link to this article: <https://doi.org/10.1080/10400419.2023.2198845>



Published online: 26 Apr 2023.



Submit your article to this journal [↗](#)



Article views: 733



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 7 View citing articles [↗](#)



Towards Game-Based Assessment of Creative Thinking

Janet Rafner^a, Qian Janice Wang^a, Miroslav Gadjacz^a, Thomas Badts^a, Brendan Baker^b, Carsten Bergenholtz^a, Michael Mose Biskjaer^a, Thomas Bui^a, Andrea Carugati^a, Matthieu de Cibeins^a, Lior Noy^c, Seyedahmad Rahimi^d, Kristian Tylén^a, Blanka Zana^a, Roger E. Beaty^b, and Jacob Sherson^{ip a}

^aAarhus University; ^bPennsylvania State University; ^cOno Academic College; ^dUniversity of Florida

ABSTRACT

For decades, researchers have struggled with measurement problems related to the *construct validity* of divergent and convergent thinking in creativity assessments. In response, some have called for battery-based approaches. Recently, digital games have emerged as a potential alternative, offering increased *scalability* and *improved ecological validity*. This article presents CREA: a new, scalable, game-based assessment suite. CREA includes *crea.tiles* and *crea.blender*, non-verbal games featuring both divergent and convergent thinking modes, as well as *crea.ideas* – the Alternative Uses Task, a standard test of divergent thinking, and *crea.logic* – a test of abstract reasoning. The novel convergent and divergent thinking game modes are constructed within the same contextual environment and with theoretically motivated differences in game-prompts to understand and generalize from the emerging elicited behaviors. In this study, 408 participants completed the CREA suite and selected validation measures, representing the largest game-based validation study to date. Both convergent and discriminant validity is demonstrated for *crea.tiles* with respect to standard tests, with correlations ranging from $r = .1$ – $.4$. Having CREA freely accessible, we aim to broaden the accessibility of creativity assessment to researchers, educators, and the general public and through this scaleup validate the rich creative behavioral patterns observed in this study.

PLAIN LANGUAGE SUMMARY

Creative thinking is an essential attribute for success in the 21st century, yet it is difficult to measure because it is complex and contextual. Traditional pen and paper tests of creative thinking are the status quo and are backed up by years of research; however, there are well-known issues with these tests such as (1) they are difficult to scale up, as they require expert human-raters to evaluate the responses; (2) many of them are verbal, which is not necessarily fair within a culturally and linguistically diverse population; and (3) there are still fundamental open questions about what they actually measure. Digital, game-based creativity assessments represent an emerging alternative to conventional creativity tests, presenting the potential for increasing scalability and improving the quality of the results. This article presents CREA: a new, scalable, suite of game-based assessments. CREA includes *crea.tiles* and *crea.blender* – non-verbal games featuring several game modes, as well as *crea.ideas* – a standard test of divergent thinking, and *crea.logic* – a test of abstract reasoning. Participants ($n=408$) were university students from North America. They completed the CREA suite and a series of validation measures. The results from the study indicate that the current, freely accessible, version of the CREA suite demonstrated sufficient validity to be used by researchers and practitioners as a scalable, creativity assessment instrument for automated figural assessment of divergent and convergent thinking.

ARTICLE HISTORY

Received May 3, 2022

Introduction

Creativity has been identified as one of the most distinguishing and valuable attributes of the workforce in the global economy (Future of Jobs Report, 2020) and is an integral part of understanding human cognition. There is a critical need for instruments that can reliably and validly assess creativity to test the efficacy of specific creativity interventions. The majority of traditional creativity instruments focus on two essential processes

of creative thinking conceptualized by Guilford (1956): divergent and convergent thinking. Divergent thinking (DT) is defined as the ability to come up with many different solutions to a prompt, whereas convergent thinking (CT) is defined as the ability to find the single best (correct) option within a space of solutions (Cropley, 2006; Guilford, 1956).

On the one hand, research and implementation of standardized, domain-agnostic tests targeting creative

processes, for instance idea generation and refinement (Bai, Duan, Kroesbergen, Leseman, & Hu, 2020; Torrance, 1976), have a long, rich history, and the quality and quantity of the instruments assessing creative constructs have substantially increased over the past 30 years (Plucker, 2022). DT measures correlate with self-reported creativity measures such as openness, creative self-concept, and the biographical inventory of creative behaviors (BICB) (Beaty & Johnson, 2021). DT measurements have been pushed in promising directions through the development of a systematic framework for scoring (Reiter-Palmon, Forthmann, & Barbot, 2019) as well as the Multi-Trial Creative Ideation framework, which reinterprets the current measurement paradigm to be more oriented toward the ideation process (Barbot 2018). Additionally, the use of semantic analysis for automated scoring of DT tasks, although limited to verbal responses (Beaty & Johnson, 2021; Johnson et al., 2022), allows for scalability of data analysis as well as catalyzed discussions on the role of semantic memory in DT tasks.

On the other hand, these same tests are currently being reevaluated due to concerns around the three fundamental aspects of any psychometric instrument: *validity*, *reliability*, and *fairness*. Firstly, only weak correlations have been established between different tests of the same construct (i.e., convergent validity) (Runco, Abdulla, Paek, Al-Jasim, & Alsuwaidi, 2016). Secondly, researchers have had difficulties in operationally disentangling the DT and CT constructs and processes (i.e., discriminant validity) (Cortes, Weinberger, Daker, & Green, 2019; Goldschmidt, 2016; Said-Metwaly, Van den Noortgate, & Kyndt, 2017). This has led to more fundamental concerns over the absence of consensus among researchers about construct definition (i.e., construct validity) (Said-Metwaly, Van den Noortgate, & Kyndt, 2017), the uncertain, real-world predictive power (i.e., ecological validity) of standard DT tasks (Plucker & Runco, 1998; Plucker, 2022; Said-Metwaly, Taylor, Camarda, & Barbot, 2022), and cultural and demographic implementation disparities, such as bias and inequity (instrument fairness) (Argulewicz, Bingenheimer, & Anderson, 1983; Becker & Cabeza, 2021). This, for instance, stems from the purely linguistic implementation of the prominent Alternative Uses Tests (AUT; Guilford, 1967) and the Remote Associates Test (RAT; Mednick, 1968) for DT and CT, respectively. Oltețeanu and Zunjani (2020) created a variant of the RAT with visual instead of verbal prompts, however, the participants' response is still verbal, which remains a barrier to bypass the linguistic hurdle and create visual, more culturally agnostic measures. These methodological

challenges are accentuated by the reliance on time-consuming and subjective scoring of creativity test responses by humans and the lack of battery-based approaches to assemble increasingly versatile and robust creativity metrics (Cortes, Weinberger, Daker, & Green, 2019; Plucker & Runco, 1998; Plucker, 2022; Runco, Abdulla, Paek, Al-Jasim, & Alsuwaidi, 2016).

Digital, game-based creativity assessments have been highlighted as a potential means to address many of the above-mentioned issues (Kaufman, 2015; Rafner et al., 2021; Shute & Rahimi, 2021; Sternberg, Glăveanu, & Kaufman, 2022). Three arguments support the usage of digital games to measure creativity. First, digital games are often largely visual rather than language-based, addressing the issue of fairness in creativity assessment with respect to language proficiency. Second, digital game environments afford more casual, natural, and immersed behavior, improving ecological validity and participant engagement. This increased potential for an intrinsically motivated flow state (Csikszentmihalyi, 1997) both offers a viable pathway to large-scale participant recruitment and can mitigate unintended psychological consequences such as test anxiety, researcher effects, and testing fatigue, which can be present in traditional testing settings (McCambridge, Witton, & Elbourne, 2014; Schram & Ule, 2019). Third, digital games can offer continuous fine-grained metrics for assessing not just creative products, but also cognitive processes using stealth assessment (an unobtrusive, ongoing assessment method for assessing hard-to-measure constructs; Rahimi & Shute, 2021; Shute, 2011). These fine-grained metrics provide a pathway to procedural investigations of creativity and lend themselves well to automated scoring of participant performance using machine learning. All of the above-listed attributes of digital game-based assessment hold potential for strengthening both ecological and construct validity.

While there are several arguments for digital game-based creativity assessments, few games currently exist for this purpose. In a recent review of game-based creativity assessments, Rafner et al. (2021) found only eleven such studies, all with low-medium numbers of participants (*Mdn* 70.5, *max* 266). Surveying the substantial differences in study design, computational and assessment approach, and theoretical foundations, the review concludes that (A) developing a robust, and non-verbal assessment for CT is one of the bigger challenges in the field; (B) the field should aim to deliberately construct digital games exploring the span from simple and stylized to complex and real-life relevant; and (C) ideally all of these separate efforts should be combined into a comprehensive suite of games.

The present study

The CREA suite

This article presents the first large-scale study of CREA, a digital game-based creativity assessment suite (Rafner, 2021) (see Figures 1, 2, and 3 below) created as a concrete response to the goals and challenges listed above. As mentioned, a psychometric evaluation instrument should be *valid*, *reliable*, and *fair*.

Within CREA, *validity* and *reliability* are pursued in a bottom-up theoretical design featuring:

- (1) DT, CT, and an open-play game mode within a carefully designed *contextually homogeneous environment* that permits introducing subtle but

nonetheless crucial differences in game prompts to more easily contrast the elicited behaviors.

- (2) A novel CT mode conceptually more closely aligned with the original CT characterization (Guilford, 1956) and the double-diamond creative design process (Design Council, 2005) than the conventionally employed RAT by featuring both *selection* and *refinement* sub-modes.
- (3) Two separate subgames, tiles and blender, with a deliberate transition between *creativity in contained* and *unrestricted state spaces* respectively, each with the structure from bullet 1.
- (4) A *battery-based approach* combining both the mini games and traditional tests, crea.ideas (measuring DT based on the Alternative Uses Task

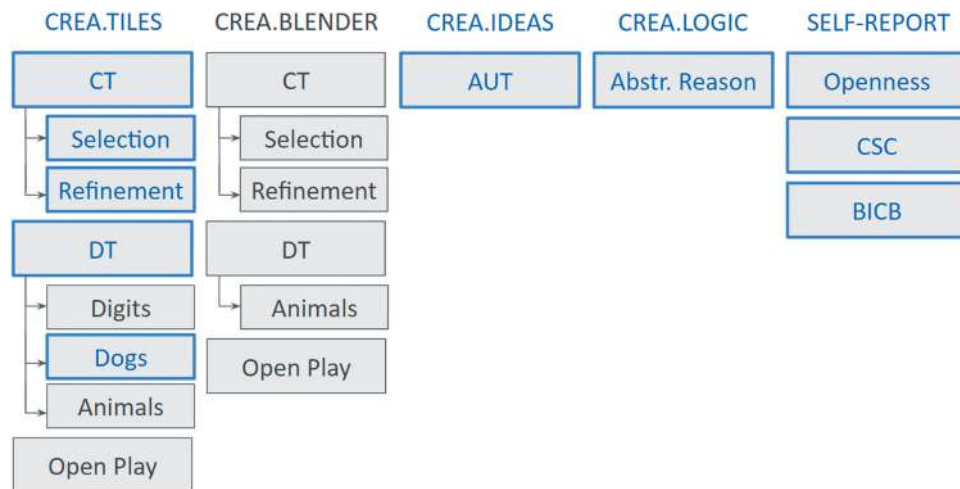


Figure 1. Overview of the structure of the CREA battery. *Note.* The abbreviated measures included are Convergent Thinking (CT), Divergent Thinking (DT), Alternative Uses Test (AUT), Abstract Reasoning (Abstr. Reason), Creative Self-Concept (CSC) and Biographical Inventory for Creative Behaviors (BICB). Measures in blue were analyzed in this article. Measures in black were not analyzed in this article. Please note that the self-reported measures are not within the CREA game ecosystem, but were administered separately via Qualtrics (Qualtrics, Provo, UT).



Figure 2. Illustration of crea.tiles. *Note.* **A:** diagram is showing which tiles can be moved and where they can be placed. Tiles marked with O can be moved to any Z locations. Tiles marked with X cannot be moved. **B:** Screenshots from tiles DT mode personal gallery. **C:** Screenshot from the selection phase of the CT mode. Here players need to choose the best starting point from the four options on the left side of the screen, in order to reach the figure on the right (larger) with the fewest moves possible.

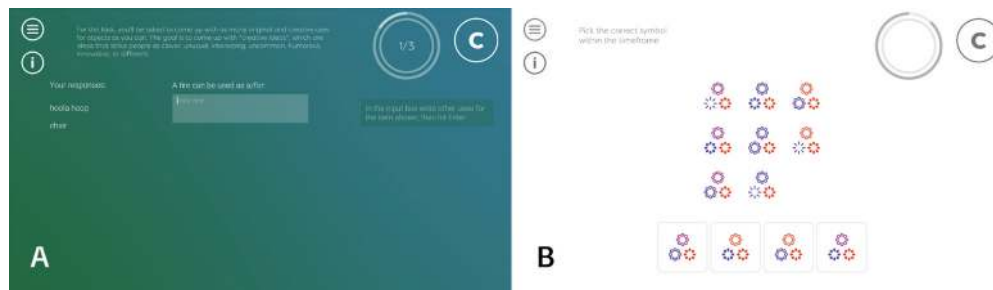


Figure 3. Screenshots of crea.ideas and crea.logic. *Note.* **A: ideas:** An example of a prompt (tire) to which players have to come up with alternative uses for. On the left side of the screen the player's submitted answers presented. **B: logic:** An example of a complex item containing more relational changes: position, color and shape. Players must select the one item from the bottom four which fills the missing square.

with automated scoring (via SemDis) (Beatty & Johnson, 2021)) and crea.logic (measuring general cognitive aspects of CT based on a variation of Raven's Progressive Matrices) in a single freely accessible package.

Fairness is pursued by combining an emphasis on *visual manipulation* and careful attention to *culturally agnostic prompts* such as the creation of dogs in the DT-mode described below.

As discussed in further detail below, the battery-based approach introduces the central development challenge and achievement of this work: to i) create a suite of interactions comprehensively covering both the authoritative representatives of established research and the theoretically motivated new games and game modes and ii) realize this within a coherent and playful unified environment. Conventional single-interaction creativity assessments are designed with sufficient time-on-task to generate robust creativity metrics such as fluency, flexibility and originality. In contrast, the central challenge when designing gamified battery testing is to consider the question: how can a state of user flow (Csikszentmihalyi, 1997) be created that allows for optimal rate of information extraction on each subtask. To our knowledge, this work represents the first battery-based application of gamification (including considerations of Mechanics, Dynamics and Aesthetics (MDA-framework; Hunnicke, LeBlanc, & Zubek, 2004) to creativity assessment. Apart from the visual game-like representation, we consider in this work explicitly the challenge of adapting prompt and timing design. Additionally, the games were designed so that the interactions were simple and the tutorial sufficient to not bias toward experienced gamers. When time-on-task is minimized, the importance of the exact formulation of the game prompts is accentuated immensely. The findings from

this work represent a crucial first step for the continuing evolution of the CREA-suite as well as for the field of gamified creativity assessment as a whole.

Here below details the theoretical motivation for the choice of games

Crea.tiles

Crea.tiles is an adaptation of the Creative Foraging Game (Hart et al., 2017, 2018, 2022; Noy et al., 2012; Rosenberg & Noy, 2022; Rozenkrantz et al., 2017). The Creative Foraging Game is a non-language-based game where participants explore the space of ten connected squares, and are asked to build shapes they find “beautiful and interesting.” Upon discovery, participants could save a particular shape in their personal gallery. Players usually play for 10–20 minutes. With a total of ~36k possible unique shapes, the product space is large from the perspective of the user but sufficiently small that, in contrast to e.g. the open-ended linguistic AUT possibility space, it is amenable to a complete mapping and expert analysis. At the same time, submitting just a few shapes will generate nearly limitless possible unique submission trajectories, yielding a rich source of behavioral patterns. This can be used to generate multi-dimensional metrics for the creative processes and outcomes. Thus, creativity, sense-making, and assessment is greatly empowered while still limiting the solution space.

In contrast to the Creative Foraging Game, which is a standalone game, crea.tiles is designed as a playful component of a suite. Thus, the single prompted, open-ended task prompt in the Creative Foraging Game is in crea.tiles adapted to several short game modes covering selectively the theoretical concepts of Convergent and Divergent thinking. The extent to which previous theoretical insights and creativity variable definitions

carry over to this broad, time-limited design is a central empirical question addressed in this work. See Methods and Materials and Discussion section for more details.

Crea.blender

In contrast to crea.tiles, crea.blender utilizes AI-supported image blending to create aesthetically refined, full resolution (512×512 pixel) images with just a few intuitive slider movements. Apart from the underlying differences in game mechanics, the task and prompt design have been retained as much as possible to preserve contextual homogeneity among the two sub-challenges. The long term goal with this game mode is to observe the changes in elicited creative behavior as the output space complexity increases compared to the same prompts within the well-contained crea.tiles mode. Such systematic investigation of divergent and convergent thinking across complexity scales is expected to yield important theoretical insights into the complex and contextual phenomenon of creativity. The AI-assistance in crea.blender is crucial to ensure that complex products can be produced sufficiently rapidly to fit within the game suite constraints. See Rafner et al., 2020 and Gajdacz et al., 2021 for exploratory studies examining user control in crea.blender. The current data acquisition was mainly used for large scale calibration of the AI-assisted task difficulty levels and in-game scoring mechanisms. First validation results of crea.blender will therefore be reported elsewhere as these calibrations are implemented into the current game-suite. However, the game mode is described here because its presence forms the crucial design considerations and time constraints for the remaining modes that encapsulates the main development challenges of calibrating and validating a time and flow optimized game-suite.

Crea.logic

Crea.logic is used for this first validation study as CT has been shown to be correlated with abstract reasoning ability (Xia, Kang, Chen, Ouyang, & Hu, 2021) and the meta-review finding that divergent thinking should be weakly correlated with intelligence ($r = .25$) (Gerwig et al., 2021). The traditional measure for CT, the RAT has been criticized for heavily assessing vocabulary, while the visual nature of crea.logic is similar to that of the CT crea.tiles and CT crea.blender modes.

Crea.ideas

Crea.ideas was chosen as a standard measure of DT, as it is a commonly used task and can be automatically scored.

Due to the comprehensive nature of this new assessment instrument and the extensive data analysis needed to unpack the results from the initial data collection ($N = 408$), the results from crea.tiles, crea.ideas, and crea.logic are reported in this article. At the time of data collection, crea.blender was in a much less mature stage than the other three components of CREA and it will thus be reported on in a future article.

Guiding questions and hypotheses

To assess the viability of the CREA suite, and in particular, crea.tiles, for creativity assessment, a study was conducted informed by the following general guiding questions:

- Is the CREA suite a valid creativity assessment battery?
- To what extent can a suite-based approach (i.e., battery of tasks, tests, and games) reveal deeper insights into the constructs underlying creativity compared to individual traditional assessment tasks?
- To what extent can scalable, digital measures of creativity supplement or possibly even replace traditional expert-assessed tasks?
- What can be learned about the relationship between creativity and general cognitive abilities (e.g., fluid intelligence (Cattell, 1963))?

The pre-registered hypotheses are as follows:

H1: As convergent and divergent thinking are conceptually distinct cognitive processes (Guilford, 1956), the performances from DT and CT modes of crea.tiles are expected to be only weakly correlated.

H2: Abstract reasoning, as measured via crea.logic, largely explains the overlap between DT and CT and is more closely aligned to performance in crea.tiles (CT) than crea.tiles (DT).

H3: Conversely, divergent thinking, as measured via the Alternative Uses Task (AUT, implemented in crea.ideas), is more closely aligned to performance in crea.tiles (DT) than in crea.tiles (CT).

H4: Divergent thinking ability, as measured by the crea.tiles (DT), is associated with self-reported measures of creative personality and behavior, including openness, creative self-concept, and biographical inventory of creative behaviors (BICB).

H5: Self-reported familiarity with games in general should not affect CREA game performance (Rahimi, 2020).

Materials and methods

Participants and procedure

This study was approved by the Institutional Review Board at Aarhus University. Participants ($n = 408$) were students at Penn State University and received optional course credit for their participation in the study. Participants gave their written informed consent before starting the study. Of the 408 in the full sample, 274 identified as female, 130 identified as male, 1 identified as non-binary, and 3 preferred not to disclose this information. Participants' ages ranged from 18 to 46 ($M = 18.95$, $SD = 2.31$).

Participants completed the battery of creativity-related assessments on their own devices and in a setting of their choice across four weeks. Participants received the following instructions "Please complete the study in a quiet place free of distractions where you can focus. The study should be completed in one sitting." In total, the activities took approximately two hours.

Game design

CREA is freely available to play in the browser here: <https://webgl.scienceathome.org/crea/>. The game was developed in Unity. See Figures 2 and 3 for screenshots from the game. In the following paragraph, one of the subgames (crea.tiles), and two gamified validation tasks (crea.ideas and crea.logic) will be described.

A note on time limits

External testing conditions such as time limits and playfulness of the testing environment are central topics which influence outcomes in DT assessment (Paek, Alabbasi, Acar, & Runco, 2021). While time limits are widely used to allow for fast and well-contained studies, some articles suggest that these constraints are actually a "creativity killer" (Said-Metwaly, Fernández-Castilla, Kyndt, & Van den Noortgate, 2020). Time limits on the DT assessments in CREA were included for the following reasons: 1) to increase beneficial "tension" in the game play, which has shown to keep players engaged in a game (MDA-framework; Hunicke, LeBlanc, & Zubek, 2004), 2) to ensure that the game could be administered in a known time frame.

Crea.tiles: CT and DT measures

Crea.tiles consists of ten connected tiles that can be moved around one at a time to form shapes (~36k possible shapes), see Figure 2A, left. The player must place a tile in an empty spot adjacent to the edge of another tile. Each play mode has several levels with explicit instructions to elicit divergent and convergent thinking. Players complete a tutorial, which introduces them to the basic game mechanics and features, then proceed through the three modes: *DT mode*, *CT mode*, and *free play*.

The **DT mode** has three levels. Players are instructed to create and save at least five different digits, dog-like, and animal-like shapes respectively within four minutes for each level. At the end of each level, players can see the shapes they have saved in their personal gallery, see Figure 2B.

The **CT mode** is designed to capture and quantify both the *option-selection* and the *selection-refinement* features of convergent thinking and thus be conceptually more closely aligned to the original CT characterization (Guilford, 1956) and the double-diamond creative design process (Design Council, 2005) than the conventionally employed RAT. Concretely, players are presented with a target figure and four possible starting figures, see Figure 2C. The objective is to select the best starting shape and rearrange it into the target shape in the fewest possible number of steps. There is always a unique optimal choice for the starting shape and the trajectory for recreating the target shape. Participants have 30 seconds to choose a starting shape and one minute to recreate the target shape. Players must solve six challenges with increasing difficulty (two easy, two medium, and two hard). The challenges were designed with an estimated difficulty based on two parameters: 1) the minimal number of steps between the optimal starting shape and the target shape; and 2) the difference between the minimal number of steps to the target from the optimal and the foil candidates, respectively. A task is deemed difficult if it takes many moves to reach the target and/or there are one or more foil candidates with a very similar distance from the target.

Based on theoretical considerations, the former two game modes, designed to disentangle DT- and CT-processes, were supplemented by an open-ended *free play* mode (more similar to the original installations of the game) where players are asked to make "the most creative shapes you can think of." When the first two modes are sufficiently validated as CT/DT measures, this construction will allow for a contextually homogeneous exploration of the importance of CT and DT in creative processes. As mentioned below, several prompt

and timing suggestions have been uncovered in this first validation work and therefore analysis of the free play modes will be postponed to future studies.

Validation tasks

crea.ideas: standard DT measure

Crea.ideas is a digital variant of the Alternate Uses Task (AUT; Guilford, 1956) and is used as a DT measure. Participants were asked to provide as many creative uses for three common object prompts (tire, brick, paper-clip), see Figure 3, left. They were instructed to “think creatively” when coming up with their ideas (Said-Metwaly, Fernández-Castilla, Kyndt, & Van den Noortgate, 2020). Participants were given two minutes to complete each prompt.

crea.logic: abstract reasoning measure

Crea.logic is a tool for assessing abstract reasoning: one’s sensitivity to relational or rule-based structure beyond the immediately or concretely observable. Abstract reasoning is thought to be related to learning, working memory, and processing speed (Markman & Gentner, 2001). Logic is implemented with the 80 items from the matrix reasoning item bank (MaRs-IB) (Chierchia et al., 2019). MaRs-IB was created and validated to serve as a free alternative to the prominent abstract reasoning test, the Raven Progressive Matrices (Raven, 2000). In MaRs-IB, players are presented with three-by-three matrices containing abstract shapes in each cell except for one, and they have to choose the missing one from a set of four options by figuring out the relationships across the provided shapes. The items

can vary in color, size, position, shape, and the combination of these, and the matrices vary in difficulty based on the amount of changing dimensions. Players have eight minutes to solve as many puzzles as they can. See Figure 3, right.

Measured variables from crea

The following performance metrics were calculated per game mode, per player.

Crea.tiles DT Mode: Out of the three DT modes (Digits, Dogs, and Animals, see Figure 2, middle), Digits was created to be the practice round as it has a very limited outcome space. DT performance was only calculated in Dogs, since “Dogs” was a reasonably constrained topic that also allowed for exploration, and a sensible deployment of “Animals” would require first a full characterization of the “Dogs” mode. To ensure data consistency between players, only the first 5 shapes submitted were used, since people were given the option to leave the game after 5 submissions. For this reason, a fluency variable for tiles-DT does not make sense in the current iteration.

Table 1 presents the theory-driven metrics for the DT mode “Dogs” along with the exploratory parameter “Duration.”

Figure 4 presents a selection of dogs that were submitted and evaluated as appropriate.

Note: In order to give the reader a basic understanding of the possible output space for the novel game-prompt “Dogs,” the first and last author reviewed all the dogs that were submitted and evaluated as appropriate. In this process they created ad hoc categories to help understand the breadth of the solution space and richness of the data set. As player intent was not recorded as

Table 1. Theory-driven metrics calculated in crea.tiles DT mode.

Frequency (originality metric)	Originality is commonly defined based on <i>statistical infrequency</i> (Runco & Acar, 2012). Therefore, shape frequency is defined as the number of times a specific shape is submitted within the five first submissions over the whole set of players. Player frequency is the average of the shape frequencies over the first five submitted shapes.
Flexibility	To calculate how distinct the submitted shapes are, flexibility is defined as the sum of the minimum distance between all pairs of shapes within the first 5 shapes submitted. This method is inspired by Runco (1986) which presented flexibility as the number of changes between categories rather than total number of categories visited.
Elaboration	The number of right angles is calculated in each given shape as an approximation of how elaborate the shape is. A straight line is the least elaborate shape containing 4 right angles. The greater the number of right angles, the higher the elaboration score. This method is inspired by the definition of elaboration as “the process of adding more information to existing, relatively simple information to create a more complex, emergent whole” (Nouri et al., 2015, p. 5).
Duration	Average time taken to create each of the first five submitted shapes. This is strictly not a metric with a theoretical link to creativity literature, however, as the aim of the study is to pursue a deep behavioral understanding of the creative processes enlisted in the game, this metric was added.
Appropriateness	Many researchers have suggested the use of a quality score (Runco, Illies, & Eisenman, 2005). Hence, a digital app was created to allow four expert raters to rate all submitted dog shapes for appropriateness. They were instructed to answer on a likert scale from 1–5 “to which extent do you think that others would recognize this as a dog on a scale” where 1= “no one” to 5 = “anyone.” High inter-rater agreement was found with Cronbach’s alpha =.87 on 983 shapes. Whereas this sample does not cover all possible submissions, it does provide sufficiently comprehensive coverage of the most commonly submitted shapes to allow for a first automated approach that can then be supplemented by occasional expert review as nonstandard shapes are submitted.

Note. Based on these initial efforts, all metrics are now readily available for direct computation in future implementations.

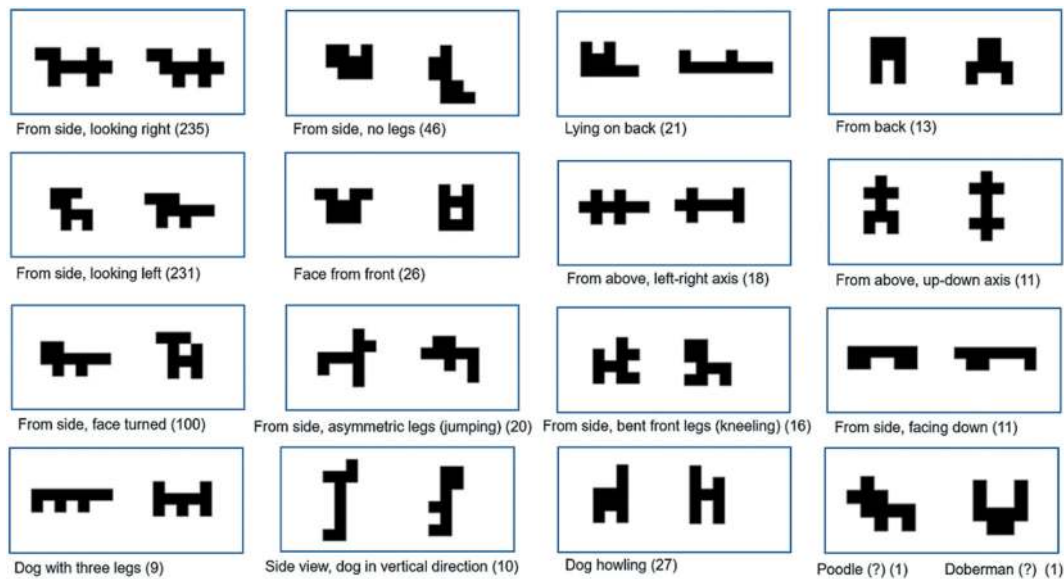


Figure 4. Examples from dog submissions. Note. Rating frequency in parenthesis.

part of the study, no quantitative analysis on these categories was conducted. See the Limitations and Future Directions section for more information.

Crea.tiles CT mode

Within the novel, two-stage procedure (choosing the correct starting shape, then moving the blocks to produce the target shape), both accuracy (correct responses) and duration for the choice stage and move stage were evaluated. Taking into account both accuracy and response time is similar to the approach of abstract reasoning assessment for MaRs-IB (Chierchia et al., 2019). To calculate a single metric, performance was represented by *standardized(choice correctness) + standardized(move correctness) - standardized(choice duration) - standardized(move duration)*.

Crea.ideas

Crea.ideas uses fluency, defined as the number of responses submitted by the participant, as well as two automated measures. First, we used an automated measure of DT based on semantic distance. Semantic distance provides a computational and objective means to quantify originality using distributional semantic modeling (Beaty & Johnson, 2021). When applied to the AUT, semantic distance captures how “far apart” objects (e.g., a tire) are from ideas (e.g., a “swing”) in high-dimensional semantic space by computing the distance between the objects and ideas (i.e., how often the object and idea co-occur together in natural language). Prior work has psychometrically

validated semantic distance, reporting positive correlations with human creativity ratings and established measures of creativity (e.g., creative achievement in the arts and sciences; Beaty & Johnson, 2021). In this study, to calculate semantic distance, the open-access platform, SemDis (<http://semdis.wlu.psu.edu/>) was used with the following text preprocessing and computational settings: remove filler and clean (which removes “stop words” and miscellaneous characters); semantic spaces = all (which computes semantic distance using all five models available on the platform); and compositional model = multiplicative (which combines the word vectors in multi-word AUT responses via multiplication; Beaty & Johnson, 2021). As the outcome variable for analysis, a composite factor score of the five semantic models (computed by SemDis) was used, averaged across the three AUT items.

A second automated measure of originality was also computed using large language models (LLMs), a class of deep learning models very effective in a broad range of natural language processing tasks due to their huge amount of parameters and training data. These features allow for good generalization capabilities when fine-tuned on downstream tasks (Wang et al., 2022), such as question answering, sentiment analysis or sentence translation. Here, GPT-3 was further trained on 27k AUT answers from previous studies (Beaty & Silvia, 2012; Beaty et al., 2018; Dumas, Organisciak, & Doherty, 2020; Hofelich Mohr, Sell, & Lindsay, 2016; Silvia et al., 2008), along with their corresponding originality human ratings. It has been shown to perform better than SemDis when it comes to matching human raters’ judgments of originality for AUT tasks (Organisciak, Acar, Dumas, & Berthiaume, 2022).

Crea.logic

Crea.logic uses a measure of abstract reasoning, calculated based on each participant's number of correct answers out of the completed items (Chierchia et al., 2019). Participants are not penalized for incorrect answers or for not answering the total set of 80 questions. See Figure 3, right for an example of matrices presented to participants.

Self-reported measures

The self-reported validation measures of creative behavior and personality were recorded on a Qualtrics survey administered by Penn State University. An online version of the McCrae and Costa (1985) Openness to Experience questionnaire was administered. This is a 10-item survey with 5-point Likert scale items (e.g., "I enjoy new ideas.") with responses from 1 (strongly disagree) to 5 (strongly agree). Furthermore, an online version of the Karwowski (2014) creative self-concept (CSC) scale was administered, since it has been found to be significantly predicted by creative performance ($r = .15$) and self-reported originality ($r = .32$) (cf. Karwowski, 2011). CSC is assessed through an 11-item questionnaire, using a scale from 1 (strongly disagree) to 5 (strongly agree) (e.g., "I think I am a creative person."). Additionally, the 34-item Biographical Inventory for Creative Behaviors (BICB) was administered in order to get an insight into participants' prior creative accomplishments, and to serve as another indicator of self-reported creativity (See Appendix). Measures of self-reported openness (ten items, averaged, *Cronbach's alpha* = 0.81), creative self-concept (11 items, averaged, *Cronbach's alpha* = 0.88), and biographical inventory of creative behaviors (34 items, 0/1, added) were combined to create a total score for openness, creative self-concept, and biographical inventory. For each self-reported measure (Openness, CSC, BICB), data was excluded from participants who failed the attention check for that particular questionnaire.

Finally, because the game performance can be influenced by familiarity with playing video games in general, following Rahimi and Shute (2021), gaming

frequency was self-reported on a scale from 1 to 5 (1 = Never play, 5 = Play almost every day).

Results

To address the research questions and hypotheses presented in this article, the results section is divided up into three subsections: Individual Measures/Descriptive Statistics, Hypothesis Testing, and Exploratory Analysis. First, individual measures are described before combining them to validate the pre-registered hypotheses. Finally, the analysis is expanded with a data-driven methodology to confirm relationships between game metrics and to explore more subtle player behaviors.

Individual measures/descriptive statistics

Crea.tiles

Correlation results among the CT and DT metrics are presented in Table 2. The correlations amongst Crea.tiles metrics revealed that appropriateness (i.e., how dog-like is each shape) is significantly and negatively correlated with frequency, flexibility, and duration, and positively correlated with elaboration. In other words, people in the study who produced more appropriate dog shapes also tended to generate more original (less frequent) and more elaborate shapes (i.e., having more right angles) which tended to be similar to each other (i.e., low flexibility) and took longer to produce. Furthermore, duration is most strongly correlated with flexibility, followed by frequency. The interpretation is that players who take longer to generate shapes also tend to produce shapes that are more different from each other and are more often produced by other players. Finally, it should be kept in mind that, since most players only submitted five shapes, producing the same shapes as many others should not necessarily be interpreted as a general lack of originality.

Crea.ideas

To assess the internal consistency of the participants' behavior while they were playing crea.ideas, pairwise correlations were run between participants' SemDis

Table 2. Pearson's correlations among crea.tiles game measures (Crea.Tiles Ct and Dt Metrics).

	Crea.tiles DT Frequency	Crea.tiles DT Flexibility	Crea.tiles DT Elaboration	Crea.tiles DT Duration	Crea.tiles DT Appropriateness
Crea.tiles CT	.088 (.22)	−0.092 (.091)	.082 (.13)	−0.14 (.010)	.090 (.098)
Crea.tiles DT Frequency	-	.41 (<.005)	−0.26 (<.005)	.18 (.015)	−0.29 (<.005)
Crea.tiles DT Flexibility	-	-	−0.35 (<.005)	.42 (<.005)	−0.40 (<.005)
Crea.tiles DT Elaboration	-	-	-	−0.02 (.67)	.72 (<.005)
Crea.tiles DT Duration	-	-	-	-	−0.14 (.011)

Note. P-values are noted in the parentheses and significant correlations are presented in bold.

scores on the different items (tire score against brick score, tire score against paperclip score, and brick score against paperclip score). Inter-item correlation between brick and tire SemDis scores is .46 ($p < .005$), and correlations between paperclip and tire and paperclip and brick are lower (respectively .12, $p = .018$ and .20, $p < .005$ for all comparisons). These results show that participants performed consistently across different items in Ideas (i.e., internal consistency; Beaty & Johnson, 2021). However, the lower correlations between paperclip and the other items indicate a lower reliability on this item. Additionally, paperclip has not been validated in previous AUT studies using semantic distance, which have recently shown differences in reliability and validity as a function of AUT item (Beaty, Johnson, Zeitlen, & Forthmann, 2022). Therefore, the paperclip was excluded from the calculation of the crea.ideas metrics.

In order to assess consistency between participants' performance in Ideas and self-reported creativity measures, a correlation was run between our metrics of Ideas (two automated scores of SemDis and originality, plus fluency) and measures of openness, CSE, and BICB. Results revealed no significant correlations of ideas SemDis score with any self-reported measures, with correlations of .04 ($p = .39$) for openness, $-.006$ ($p = .92$) for CSC, and .007 ($p = .91$) for BICB, consistent with some prior work showing low correlations between AUT semantic distance and self-reported scales (Beaty & Johnson, 2021; Beaty, Johnson, Zeitlen, & Forthmann, 2022). Similar results were obtained with the machine-learning based originality scoring (openness: $r = .047$, $p = .40$; CSC: $r = -.013$, $p = .82$; BICB: $r = .032$, $p = .56$). In contrast, Ideas fluency, defined as the number of solutions provided by the participant, had a significant positive correlation with BICB ($r = .17$, $p = .002$), but not for openness ($r = .028$, $p = .62$) or CSC ($r = .09$, $p = .11$).

Crea.logic

The mean number of correct answers was 53% (SD = 16%), and the median response time on correct items was 4.91 s (interquartile range (IQR) = 7.04 s). On

average, participants completed 54.7 unique items (SD = 17.8) within 8 minutes. Timing out occurred in 1.1% of items. A skewness of .28 and a kurtosis of 2.20 suggest that participants' scores fell reasonably close to a normal distribution. Compared to the original Mars-IB validation paper (Chierchia et al., 2019), participants were faster and more inaccurate. However, the results found in the present study were in line with follow up work using the same instrument (Nussenbaum, Scheuplein, Phaneuf, Evans, & Hartley, 2020).

Self-report scales

The three self-report measures of openness, CSC, and BICB showed significant positive correlations among them, as expected. Openness and CSC had the highest correlation ($r = .69$, $p < .005$), followed by openness and BICB ($r = .30$, $p < .005$), and CSC and BICB ($r = .30$, $p < .005$). The results show that participants who self-report having a high creative self concept also self-report having a high tendency toward openness to experiences and self-report high scores on past creative achievements.

Hypothesis testing

The correlations applied to test hypotheses H1-H4 are reported in Table 3. As per standard practice, outliers (defined as data points at least three standard deviations from the mean) were removed before running the correlation calculations.

According to H1, it was predicted that the DT and CT mode of Crea.tiles would capture different cognitive processes. In fact, according to Table 3, the only significant correlation between DT and CT is in duration time ($r = -.14$, $p = .010$). As the CT metric is negatively weighted toward the time taken to complete each trial, this correlation can be explained by consistencies in player reaction time across the DT and CT modes in crea.tiles. There are no other significant correlations between DT and CT metrics, suggesting that the DT and CT modes of Crea.tiles are associated with different cognitive processes.

Table 3. Correlations among game measures (Crea.Tiles CT and DT Metrics), validation tests, (ideas, logic) and self-reported measures, with p-values in parentheses.

	Crea.tiles CT	Crea.tiles DT Frequency	Crea.tiles DT Flexibility	Crea.tiles DT Elaboration	Crea.tiles DT Duration	Crea.tiles DT Appropriateness
Crea.ideas fluency	.04(.45)	.044(.55)	.067 (.23)	-.08 (.15)	-.014 (.012)	-.087 (.12)
Crea.ideas semdis	-.09 (.09)	.11 (.12)	.017 (<.005)	-.010 (.85)	.014 (.016)	0.0035 (.95)
Crea.ideas open scoring	0.032 (.54)	0.034(.65)	.012 (.026)	0.021 (.70)	0.047 (.41)	0.037 (.51)
Crea.logic	.014 (.008)	0.12 (.10)	0.077 (.18)	0.087 (.13)	-.077 (.19)	0.00 (.93)
Openness	-.014 (.012)	-.010 (.16)	-.014 (.81)	0.047 (.42)	0.006 (.92)	0.052 (.36)
BIC	-.041 (.45)	0.026 (.73)	-.034 (.61)	0.08 (.16)	-.011 (.062)	0.078 (.17)
CSC	-.006 (.91)	-.05 (.51)	-.029 (.61)	0.042 (.46)	-.044 (.45)	0.083 (.15)

According to H2, it was predicted that abstract reasoning, as measured by the validation task *crea.logic*, should be more closely aligned to performance in the CT mode of *crea.tiles* than the DT mode. As predicted, *crea.tiles* (CT) has a significant correlation with *crea.logic* ($r = .14$, $p = .008$), but not with any of the *crea.tiles* (DT) metrics (Table 3). In other words, players who performed better in *crea.logic* (i.e., demonstrated better abstract reasoning performance), also tended to be better at visualizing and planning tiles to match the target shape.

According to H3, it was predicted that DT, as measured by the validation task *crea.ideas*, would conversely be more closely aligned to performance in *crea.tiles* (DT) compared to *crea.tiles* (CT). This is confirmed in Table 3, where several metrics of *crea.tiles* (DT) were found to have significant correlations with the *Crea.ideas* SemDis score ($r = .17$, $p < .005$ between DT flexibility and ideas $r = .14$, $p = .016$ between DT duration and ideas). Using the novel machine learning based originality scoring for Ideas yielded similar results between DT flexibility and Ideas ($r = .12$, $p = .026$). In other words, those players who performed better in *crea.ideas* also tended to produce shapes that are more distinct from each other. Additionally there were no significant correlations between *crea.ideas* and *crea.tiles* (CT), further supporting H3.

According to H4, it was predicted that divergent thinking ability as measured by the DT mode of *crea.tiles* would correlate with self-reported measures of openness, creative self-concept, and biographical inventory of creative behaviors. However, no significant correlations were found with any self-reported measures. As stated earlier, however, no correlations were observed of self-reported measures with *crea.ideas* (AUT), an established measure of divergent thinking, consistent with past work reporting mixed correlations with self-report measures of creative personality and behavior (Beaty & Johnson, 2021).

According to H5, it was predicted that self-reported familiarity with games, in general, would not be associated with CREA game performance. On a scale of 1–5 (1 = Never play, 5 = Play almost every day), the mean average gaming frequency from the participant sample was 2.2 (SD = 1.2). There were no significant correlations between gaming frequency and any CREA metrics (see Appendix).

Exploratory analysis

Data-driven analysis

In addition to the hypothesis-driven analysis on the pre-defined aggregate metrics described above, an

explorative data-driven analysis taking into account also the individual submission pathways was performed. For each player, a DT(CT)-array was constructed consisting of the creativity metrics mentioned above for each of the 5(6) submissions. These were then treated as variables for a multi-dimensional linear-model regression analysis to predict the individual *crea.ideas* and *crea.logic* outcomes. To prevent overfitting, elastic-net modeling (Pedersen et al., 2020) was chosen, which combines ridge and lasso regression by adding regularization terms to the loss function. For *crea.tiles* CT, the only significant correlation was for *crea.logic*, yielding a medium strength with high significance: $r = .37$, $r_mean = .25$, $p < .005$. For *crea.tiles* DT, statistically significant correlations were found with *crea.ideas* ($r = .39$, $r = .15$, $p = .02$) (see Appendix).

Uncovering player behavior patterns

Beyond hypothesis testing, we conducted exploratory analyses to investigate mental processes behind the participants' creative behavior. The goal of uncovering player behavior patterns is to better understand individual differences in creative propensities and identify more nuanced indicators of creative behavior. Toward this end, we include duration in the correlation analysis and observe an exploratory correlation between *crea.ideas* performance and DT duration, which lends support to incorporating this variable into theoretical considerations in future studies. Additionally, the question arose if distinct groups of behavior patterns in creative shape generation can be identified. For this, a cluster analysis was performed on each user's game metrics for *Crea.tiles* (DT). More specifically, the five DT metrics of frequency, elaboration, appropriateness, duration, and per-shape flexibility (defined as the number of minimum moves from the previous shape to the current shape) were taken for each of the first five shapes produced by a player, for a total of 25 metrics per player.

After performing agglomerative hierarchical clustering (AHC), a visual inspection of the dendrogram identified four main clusters of players (see Appendix for dendrogram) (Saas, Guitart, & Periañez, 2016). Mean values of game metrics can be found in Table 4 below. One-way analyses of variance (ANOVA) revealed that these game metrics values are significantly different based on cluster membership ($F_{\text{Frequency}}(3,340) = 99.21$, $F_{\text{Flexibility}}(3,340) = 59.03$, $F_{\text{Elaboration}}(3,340) = 85.52$, $F_{\text{Duration}}(3,340) = 18.41$, $F_{\text{Appropriateness}}(3,340) = 172.44$, $p < .005$ for all tests).

The largest cluster, Cluster 1, is characterized by players who made the most diverse shapes, which consequently took the longest time. Interestingly, players in Cluster 1: *Common Creators* also produced very

Table 4. Mean values for DT metrics for participants in each crea.tiles DT cluster.

Measures		Cluster 1 (N = 136)	Cluster 2 (N = 116)	Cluster 3 (N = 64)	Cluster 4 (N = 28)
Crea.tiles DT metrics	Frequency	26 (14) _A	8 (10) _B	6 (5) _B	34 (4) _C
	Flexibility (per shape)	3.4 (0.7) _A	2.5 (0.5) _B	2.5 (0.9) _B	3.2 (0.7) _A
	Elaboration	11.3 (2.0) _A	14.3 (1.4) _B	10.4 (1.9) _C	11.7 (2.0) _A
	Duration	25 (11) _A	16 (11) _B	16 (10) _B	18 (6.6) _B
	Appropriateness	1.3 (0.6) _A	2.6 (0.6) _B	0.9 (0.4) _C	1.5 (0.8) _A

Note: key-HSD post-hoc test.

frequent shapes, which implies that these players explored common but different dog shapes (e.g., view from the left, from the front, from the back, etc.). Cluster 2: *Highly Creatives*, at almost the size as Cluster 1, is characterized by players who produced the most elaborate, appropriate, and very original shapes relatively quickly, implying that these may be the most “creative” players. Cluster 3: *Low-effort Creators* contains players who quickly produced very common shapes with low appropriateness and low elaboration; in other words, a group of players with low motivation. The smallest cluster, Cluster 4: *Speedy Common Creators*, is characterized by players who made diverse but commonly produced shapes, similar to Cluster 1, except at a faster speed. These could be players who quickly thought of different canonical dog shapes, pointing to fast problem-solving abilities (which is shown in Cluster 4’s relatively high Crea.logic scores). Figure 4 illustrates examples of shapes produced by players of each cluster. An interactive online tool was created to better visualize these clusters of players and the shapes they have produced (see Appendix).

Discussion

Summary of results and significance

Convergent and divergent validity of crea.tiles

With 408 participants, this study presents the largest game-based creativity validation to-date (Rafner et al., 2021). It furthermore constitutes a first-time validation of a novel gamified creativity assessment suite with contextually homogenous tests of convergent and divergent thinking. In particular, the context for both tasks was the visual creation of meaningful shapes based on ten manipulable tiles. This environment was chosen because it supports both a prompt-based DT task, very similar to the AUT task (create different dog shapes), as well as a goal-oriented CT task more closely aligned with the original theoretical underpinnings of convergent thinking than the traditional RAT task by including both a selection and a refinement mode. The primary question posed in this study is whether or not the contextually similar but subtly different tiles (DT) and tiles (CT) game modes elicit the theoretically predicted behaviors. In terms of internal game-metric

correlations, the CT-metric exhibits no significant correlations with the DT-metrics (frequency, flexibility, elaboration, and appropriateness) and only a very weak significant correlation with DT-duration, suggesting that the DT and CT modes of Crea.tiles are associated with different cognitive processes (hypothesis H1). As theoretically expected the individual DT-metrics exhibit significant correlations ranging from weak to strong internally, as these metrics are known to be overlapping to still distinct. Crea.tiles (CT) shows convergent validity with crea.logic, a prominent test of abstract reasoning ability and, just as importantly, shows no correlation with crea.ideas, a standard test of DT. Conversely, two metrics of the crea.tiles DT mode are correlated with crea.ideas and only a weakly significant correlation with crea.logic was observed. This constitutes support for hypotheses H2 and H3.

Unexpectedly, self-reported measures of openness, creative self-concept, and creative activities were not correlated with any measures of divergent thinking, including both crea.tiles (DT) and crea.ideas. Finally, no correlations were found between general game experience and any CREA game metrics. This constitutes support for hypothesis H5 and implies that CREA is suitable for a wide population of players and is not biased against those with limited gaming experience.

Player behavior

Although the metrics are theoretically motivated, they are newly defined for the CREA suite. To ensure that the quality of the observed correlations is not an artifact of the particular operationalization of the concepts, an explorative data-driven analysis was conducted. This analysis takes into account the individual submission pathways rather than just the average metrics that have been performed. This analysis supports the hypothesis-driven findings. Overall, highly significant correlations in the range $r = 0.1$ – 0.3 have been identified with the standard tests and the obtained parameter weights are overall fairly consistent with the pre-registered parameter definitions. The combined analysis clearly documents that game performance from the CT and DT modes of crea.tiles are indeed associated with non-orthogonal but clearly distinct cognitive abilities as predicted both theoretically and from past research.

Secondly, preliminary player clustering analysis revealed different patterns of player behavior, which invites a further analysis of individual problem-solving types and their importance for further design and prompt evolutions, see Figure 6. Reflecting on these results, it can be seen that these player behavioral clusters embed both player traits (e.g. those in Cluster 3 appear to lack motivation as seen in the simplistic and non-appropriate shapes produced, Figure 5) and creative propensities (e.g., those in Cluster 2 producing both original and elaborate shapes). Beyond this, future studies should focus on extracting more nuanced player behavioral traits to construct robust creativity predictors in order to harvest the full fruits of this game-based design.

Limitations and future development

This study has some notable conceptual and logistical limitations that can inform future development.

In the DT mode of crea.tiles used in the study reported on in this article, players were allowed to create

as many shapes as they wanted within two minutes, but they were able to progress to the next game mode as soon as they submitted five shapes. This was intended to give the participants a sense of game-like freedom by allowing them to continue through the game if they ran out of ideas rather than having to wait for time to run out. However, this feature encouraged disparate user behavior, where some players devoted much more time on the task and produced over 20 shapes, while most others stopped after submitting five shapes. Consequently, to ensure consistency among players, only the first five shapes submitted by players were included in the data analysis. Moreover, no globally meaningful measures of fluency and flexibility could be captured since players could stop after five shapes. With only five submissions it is not meaningful to expect the presence or absence of category switches in a player submission record to accurately predict the degree of category switches in a more conventional flexibility study with a longer submission record. Therefore, the “flexibility” metric introduced here should rather be taken as an in-game behavioral



Figure 5. Three very similar dog shapes. *Note.* Frequency of shapes from left to right: 2, 5, 3.

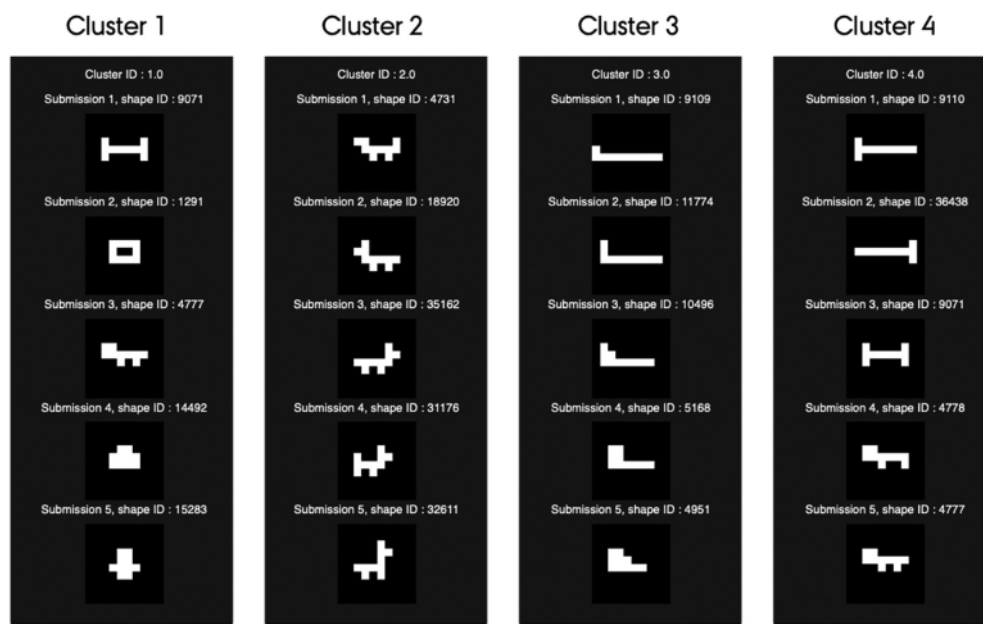


Figure 6. Visualization of shapes produced by players in different behavioral clusters. *Note.* Players in Cluster 1 tended to produce familiar shapes, Cluster 2 players produced the most original and elaborate shapes, Cluster 3 players produced simple, non-dog-like shapes, and players in Cluster 4 produced the most common shapes very quickly.

indicator than a directly generalizable creativity metric. In the future, the instruction will be changed to a time-limited, but not submission-limited, model and re-validated. In particular, the duration for the “dogs” DT mode will be extended from two to four minutes and the duration of the introductory “digits” challenges will be reduced to one minute in order to ensure more time for the generation of robust fluency and flexibility metrics. Finally, the existing instructions asked players to create “as many dog-like shapes” as possible, without any explicit prompts to solicit creative behavior, which is an important instruction in creativity assessment (Said-Metwaly, Fernández-Castilla, Kyndt, & Van den Noortgate, 2020). This will also be adapted and re-validated in future studies.

The logistical limitations described above may have contributed to the lack of correlations and low effect sizes between the DT measures from Crea.ideas and Crea.tiles and the self-reported measures (see Table 3). However, there is a general problem in the literature that studies often fail to observe strong correlations between different metrics of creativity (Batey, Furnham, & Safiullina, 2010; Carson, Peterson, & Higgins, 2005; Dollinger, Urban, & James, 2004; Furnham, Batey, Anand, & Manfield, 2008; Haase, Hoff, Hanel, & Innes-Ker, 2018), so the results presented in this article are not an exception.

A primary aim of this experiment was to determine discriminant validity between the crea.tiles CT and DT tasks. This was done using a test of abstract reasoning, the MaRs-IB implemented in crea.logic, instead of the commonly used Remote Associates Test (RAT; Mednick, 1968), since RAT is known to involve mixtures of both CT and DT (Cortes, Weinberger, Daker, & Green, 2019). However, it is a potential limitation that a test of abstract reasoning was used as a validation for convergent thinking, instead of the more canonical RAT. That said, as the crea.tiles game is a visual and figural game that does not involve verbal processing, a figural test of abstract reasoning may serve as a better validation than the verbally based RAT.

The algorithmic metric of originality used for hypothesis testing identifies the frequency of unique shapes created by users (e.g. how many other people created the exact same shape). However, upon inspecting the submitted figures, it became clear that many of them were distinct combinations of tiles, yet very similar in meaning (Figure 5). The visual similarity between close figures was not captured by the frequency metric, or any other measure. This observation was also made in the original Creative Foraging Game work by Hart et al. (2017), in which human rating of similarity within an exploitation phase (a phase of several, densely saved

gallery shapes) was used to show the validity of segmenting the trajectory of shapes into exploration and exploitation phases. The main limitation inhibiting a deeper analysis of this aspect was that there was no indication of the creator’s intent. The lack of creator intent could in principle be alleviated by allowing for participant tagging of their submissions. This tagging would however turn the purely visual task back into a linguistic task. Furthermore, even allowing for one-word tagging would explode the space of possible submissions and reintroduce the conceptual difficulties of obtaining a detailed understanding of the individual submissions and their relationships that was the main design goal of the crea.tiles game. Therefore, future studies could instead employ post-submission crowd-evaluation and hybrid intelligence (i.e. optimal synergies between humans and AI, Rafner et al., 2022) to produce a meaningful shape clustering and identification algorithm. The initial author-created ad hoc categories will be used as a point of departure for the next iteration of crea.tiles which will include a user-self-clustering functionality.

Additionally, several experimental manipulations could be conducted to further support this article’s findings. For example, future studies could involve changing the order of the games to ensure that the findings presented in this article are not an artifact of any particular game play sequence. Additionally, future studies could allow a subset of users to have unlimited time to complete the DT tasks to check that the time limit does not set an unnatural constraint on player behavior (Said-Metwaly, Fernández-Castilla, Kyndt, & Van den Noortgate, 2020). Furthermore, future studies should employ test-retest methods to ensure that the testing instrument is reliable (Groth-Marnat, 2009).

Lastly, from a data collection perspective, this study is limited by a very homogenous sample of US university students from the same institute, who predominantly identify as women, and are between the ages of 18–20. This weakens the generalizability of the study to other populations and cultures, which future studies should address. A current constraint is that the SemDis algorithm only works on English responses. However, the great advances in large language models (LLMs) over the past years make it increasingly likely that SemDis can be applied to multi-language responses in the near future (Scao et al., 2022).

In conclusion, the results from the study indicate that the current version of the CREA suite demonstrates sufficient validity to be used by researchers and practitioners as a scalable, creativity assessment instrument for automated figural assessment of DT and CT. Future development of the tool is planned, with rigorous

psychometric testing to be conducted before new versions are released. This work is expected to be useful for researchers and practitioners in a range of contexts, including large-scale assessments of students or employees and lab studies in which conditions make it difficult for standardized tests to be administered (e.g., microdosing studies).

Open access

The study is pre-registered on OSF at <https://osf.io/e2r3a>.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the Carlsbergfondet; National Science Foundation [DRL-1920653]; Novo Nordisk Fonden

ORCID

Jacob Sherson  <http://orcid.org/0000-0001-6048-587X>

References

- Argulewicz, E. N., Bingenheimer, L. T., & Anderson, C. C. (1983). Concurrent validity of the PPVT-R for Anglo-American and Mexican-American students. *Journal of Psychoeducational Assessment*, 1(2), 163–167. doi:10.1177/073428298300100207
- Bai, H., Duan, H., Kroesbergen, E. H., Leseman, P. P., & Hu, W. (2020). The benefits of the learn to think program for preschoolers' creativity: An explorative study. *The Journal of Creative Behavior*, 54(3), 699–711. doi:10.1002/jocb.404
- Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in Psychology*, 9, 2529.
- Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge and personality as predictors of creativity. *Learning and Individual Differences*, 20(5), 532–535. doi:10.1016/j.lindif.2010.04.008
- 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), 1–16. doi:10.1080/10400419.2022.2025720
- Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q. . . . 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. doi: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. doi:10.1037/a0029171
- Becker, M., & Cabeza, R. (2021). *Assessing creativity independently of language: A normed language-independent remote associate task (LI-RAT)* (pp. 1–18). Behavior Research Methods.
- Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, 17(1), 37–50. doi:10.1207/s15326934crj1701_4
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1. doi:10.1037/h0046743
- Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S. J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, 6(10), 190232. doi:10.1098/rsos.190232
- Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, 27, 90–93. doi:10.1016/j.cobeha.2018.09.017
- Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, 18(3), 391–404. doi:10.1207/s15326934crj1803_13
- Csikszentmihalyi, M. (1997). Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39.
- Design Council UK. (2005). The Design Process. <https://www.designcouncil.org.uk/news-opinion/double-diamond-universally-accepted-depiction-design-process>
- Dollinger, S. J., Urban, K. K., & James, T. A. (2004). Creativity and openness: Further validation of two creative product measures. *Creativity Research Journal*, 16(1), 35–47. doi:10.1207/s15326934crj1601_4
- Dumas, D., Organisciak, P., & Doherty, M. D. (2020). 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
- Furnham, A., Batey, M., Anand, K., & Manfield, J. (2008). Personality, hypomania, intelligence and creativity. *Personality and Individual Differences*, 44(5), 1060–1069. doi:10.1016/j.paid.2007.10.035
- Gajdacz, M., Rafner, J., Langsford, S., Hjorth, A., Bergenholtz, C., Biskjaer, M. M., . . . & Sherson, J. F. (2021). CREA. blender: A GAN Based Casual Creator for Creativity Assessment. In ICCG (pp. 405–409).
- Gerwig, A., Miroshnik, K., Forthmann, B., Benedek, M., Karwowski, M., & Holling, H. (2021). The relationship between intelligence and divergent thinking—a meta-analytic update. *Journal of Intelligence*, 9(2), 23. doi:10.3390/jintelligence9020023
- Goldschmidt, G. (2016). Linkographic evidence for concurrent divergent and convergent thinking in creative design. *Creativity Research Journal*, 28(2), 115–122. doi:10.1080/10400419.2016.1162497

- Groth-Marnat, G. (2009). *Handbook of psychological assessment*. John Wiley & Sons.
- Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, 53(4), 267. doi:10.1037/h0040755
- Guilford, J. P. (1967). *The nature of human intelligence*.
- Haase, J., Hoff, E. V., Hanel, P. H., & Innes-Ker, Å. (2018). A meta-analysis of the relation between creative self-efficacy and different creativity measurements. *Creativity Research Journal*, 30(1), 1–16. doi:10.1080/10400419.2018.1411436
- Hart, Y., Goldberg, H., Striem-Amit, E., Mayo, A. E., Noy, L., & Alon, U. (2018). Creative exploration as a scale-invariant search on a meaning landscape. *Nature Communications*, 9(1), 1–11. doi:10.1038/s41467-018-07715-8
- Hart, Y., Kosoy, E., Liquin, E. G., Leonard, J. A., Mackey, A. P., & Gopnik, A. (2022). The development of creative search strategies. *Cognition*, 225, 105102. doi:10.1016/j.cognition.2022.105102
- Hart, Y., Mayo, A. E., Mayo, R., Rozenkrantz, L., Tendler, A., Alon, U., & Noy, L. (2017). Creative foraging: An experimental paradigm for studying exploration and discovery. *PLoS One*, 12(8), e0182133. doi:10.1371/journal.pone.0182133
- Hofelich Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, 34(3), 347–359. doi:10.1177/0894439315588736
- Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *Proceedings of the AAAI Workshop on Challenges in Game AI*, 4(1), July:1722. <https://www.aaai.org/Papers/Workshops/2004/WS-04-04/WS04-04-001.pdf>
- Johnson, D. R., Kaufman, J. C., Baker, B. S., Patterson, J. D., Barbot, B., Green, A. E., and Beaty, R. E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 1–34. doi:10.3758/s13428-022-01986-2
- Karwowski, M. (2011). It doesn't hurt to ask . . . but sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts*, 5(2), 154. doi:10.1037/a0021427
- Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1), 62–70. doi:10.1037/a0034898
- Kaufman, J. C. (2015). Why creativity isn't in IQ tests, why it matters, and why it won't change anytime soon probably. *Journal of Intelligence*, 3(3), 59–72. doi:10.3390/jintelligence3030059
- Markman, A. B., & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, 52(1), 223–247. doi:10.1146/annurev.psych.52.1.223
- McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, 67(3), 267–277. doi:10.1016/j.jclinepi.2013.08.015
- McCrae, R. R., & Costa, P. T., Jr. (1985). Openness to experience. *Perspectives in Personality*, 1, 145–172.
- Mednick, S. A. (1968). The remote associates test. *The Journal of Creative Behavior*, 2(3), 213–214. doi:10.1002/j.2162-6057.1968.tb00104.x
- Nouri, R., Erez, M., Lee, C., Liang, J., Bannister, B. D., & Chiu, W. (2015). Social context: Key to understanding culture's effects on creativity. *Journal of Organizational Behavior*, 36(7), 899–918. doi:10.1002/job.1923
- Noy, L., Hart, Y., Andrew, N., Ramote, O., Mayo, A. E., & Alon, U. (2012). A quantitative study of creative leaps. *ICCC* (pp. 72–76).
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology*, 6(1). doi:10.1525/collabra.17213
- Oltețeanu, A. -M., & Zunjani, F. H. (2020). A visual remote associates test and its validation. *Frontiers in Psychology*, 11, 26. doi:10.3389/fpsyg.2020.00026
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2022). *Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models*. doi:10.13140/RG.2.2.32393.31840
- Paek, S. H., Alabbasi, A. M. A., Acar, S., & Runco, M. A. (2021). Is more time better for divergent thinking? A meta-analysis of the time-on-task effect on divergent thinking. *Thinking Skills and Creativity*, 41, 100894. doi:10.1016/j.tsc.2021.100894
- Pedersen, M. K., Díaz, C. M. C., Alba-Marrugo, M. A., Amidi, A., Basaiawmoit, R. V., Bergenholtz, C., & Sherson, J. F. (2020). Cognitive abilities in the wild: Population-scale game-based cognitive assessment (Version posted online). *arXiv preprint arXiv:2009.05274*. Retrieved April 20, 2021.
- Plucker, J. A. (2022). The patient is thriving! current issues, recent advances, and future directions in creativity assessment. *Creativity Research Journal*, 1–13. doi:10.1080/10400419.2022.2110415
- Plucker, J. A., & Runco, M. A. (1998). The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. *Roeper Review*, 21(1), 36–39. doi:10.1080/02783199809553924
- Rafner, J., Hjorth, A., Risi, S., Philipsen, L., Dumas, C., Biskjær, M. M., . . . & Sherson, J. (2020, November). CREA. Blender: a neural network-based image generation game to assess creativity. In Extended abstracts of the 2020 annual symposium on computer-human interaction in play (pp. 340–344).
- Rafner, J. (2021, June). Creativity assessment games and crowdsourcing. In *Creativity and Cognition* (pp. 1–5)
- Rafner, J., Biskjær, M. M., Zana, B., Langsford, S., Bergenholtz, C., Rahimi, S., & Sherson, J. (2022). Digital games for creativity assessment: Strengths, weaknesses and opportunities. *Creativity Research Journal*, 34(1), 28–54.
- Rafner, J., Gajdacz, M., Kragh, G., Hjorth, A., Gander, A., Palfi, B., & Sherson, J. (2022). Mapping citizen science through the lens of human-centered AI. *Human Computation*, 9(1), 66–95. doi:10.15346/hc.v9i1.133
- Rahimi, S. (2020). Inspire, instruct, or both? Game-based assessment and support of creativity (Dissertation No. 28153346) [Doctoral dissertation, The Florida State University]. ProQuest Dissertations Publishing.
- Rahimi, S., & Shute, V. J. (2021). First inspire, then instruct to improve students' creativity. *Computers & Education*, 174, 1–27. doi:10.1016/j.compedu.2021.104312

- Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1), 1–48. doi:10.1006/cogp.1999.0735
- 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. doi:10.1037/aca0000227
- Rosenberg, G., & Noy, T. (2022). Social interaction modulates collective creativity, Proceedings from the Annual Meeting of the Cognitive Science Society, CA: Toronto
- Rozenkrantz, L., Mayo, A. E., Ilan, T., Hart, Y., Noy, L., Alon, U., & Manalo, E. (2017). Placebo can enhance creativity. *PLoS One*, 12(9), e0182466. doi:10.1371/journal.pone.0182466
- Runco, M. A. (1986). Flexibility and originality in children's divergent thinking. *The Journal of Psychology*, 120(4), 345–352. doi:10.1080/00223980.1986.9712632
- Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best? *Creativity Theories—research-Applications*, 3(1), 4–18. doi:10.1515/ctra-2016-0001
- 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
- Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *The Journal of Creative Behavior*, 39(2), 137–148. doi:10.1002/j.2162-6057.2005.tb01255.x
- Saas, A., Guitart, A., & Periañez, Á. (2016). Discovering playing patterns: Time series clustering of free-to-play game data. *IEEE Conference on Computational Intelligence and Games (CIG)*, 2016, pp. 1–8, doi: 10.1109/CIG.2016.7860442.
- Said-Metwaly, S., Fernández-Castilla, B., Kyndt, E., & Van den Noortgate, W. (2020). Testing conditions and creative performance: Meta-analyses of the impact of time limits and instructions. *Psychology of Aesthetics, Creativity, and the Arts*, 14(1), 15. doi:10.1037/aca0000244
- Said-Metwaly, S., Taylor, C. L., Camarda, A., & Barbot, B. (2022). Divergent thinking and creative achievement—how strong is the link? An updated meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. doi:10.1037/aca0000507
- Said-Metwaly, S., Van den Noortgate, W., & Kyndt, E. (2017). Methodological issues in measuring creativity: A systematic literature review. *Creativity Theories—research-Applications*, 4(2), 276–301. doi:10.1515/ctra-2017-0014
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Schram, A., & Ule, A. (Eds.). (2019). *Handbook of research methods and applications in experimental economics*.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.
- Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, 116, 106647. doi:10.1016/j.chb.2020.106647
- Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I. . . . 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. doi:10.1037/1931-3896.2.2.68
- Sternberg, R. J., Glăveanu, V., & Kaufman, J. C. (2022). In quest of creativity: Three paths toward an Elusive Grail. *Creativity Research Journal*, 1–21. doi:10.1080/10400419.2022.2107299
- Torrance, E. P. (1976). *Creativity testing in education*. Creative Child & Adult Quarterly.
- Wang, T., Roberts, A., Hesslow, D., Scao, T. L., Chung, H. W., Beltagy, I. Raffel, C. (2022). What language model architecture and pretraining objective work best for zero-shot generalization? *arXiv preprint arXiv:2204.05832*.
- World Economic Forum. (2020). *Future of jobs report*. <https://www.weforum.org/reports/the-future-of-jobs-report-2020>
- Xia, T., Kang, M., Chen, M., Ouyang, J., & Hu, F. (2021). Design training and creativity: Students develop stronger divergent but not convergent thinking. *Frontiers in Psychology*, 12. doi:10.3389/fpsyg.2021.695002

Appendix

Crea.Tiles instructions.

Mode	Level	Instructions	Feedback	Timing
Tutorial	Tutorial	Complete the onboarding		untimed
CT	Challenge – Selection	Select a shape from the four small shapes on the left from which you think it will take the least amount of moves to recreate the target shape on the right. You have 30 seconds to pick the best shape and 1 min to recreate the target	Failure: Time's up! You didn't match the target image in time. Click continue to get a new set Success: [Moving on to creation]	30 sec
	Challenge – Creation	Recreate the target shape you see on the right. You have one minute	Failure: Time's up! You didn't match the target image in time. Click continue to get a new set. Success (not best shape): You did not choose the best shape. Your moves: [number] Min moves: [smaller number] Success (best shape): You chose the best shape! Your moves: [number] Min moves: [number] Well done!	1 mn
DT	Creation-Digits	Create and save as many digit-like shapes as possible. You have 4 minutes. (min 5)	Success: [Displays saved creations of the user] Success (timeout)*: Time's up. Click CONTINUE to finish. Failure (timeout): [Stays on page]	4mn
	Creation – Dogs	Create and save as many dog-like shapes as possible. You have 4 minutes. (min 5)	Success: [Displays saved creations of the user] Success (timeout): Time's up. Click CONTINUE to finish. Failure (timeout): [Stays on page]	4mn
	Creation – Animals	Create and save as many animal-like (e.g., birds, insects or reptiles) as possible. You have 4 minutes. (min 5)	Success: [Displays saved creations of the user] Success (timeout): Time's up. Click CONTINUE to finish. Failure (timeout): [Stays on page]	4mn
Open Play	Creation	Create the most creative shapes you can think of. You have 4 minutes. (min 5)	Success: [I'M ALL OUT OF IDEAS button appears after 5 - brings to display of creations of the user] Pick your 3 favorite shapes. Success (timeout): Time's up. Click CONTINUE to finish.	4 mn

*Success (timeout) = user has created at least 5 images, then timed out.

*Failure (timeout) user fails to create any images.

Crea Metrics

Descriptive statistics of the Crea metrics used in the study. The value range for each measure is presented in parenthesis in column one.

Metric	Mean	SD
Crea.tiles CT [-15-5]	.0	1.89
Crea.tiles DT Frequency [1-336]	16.89	35.83
Crea.tiles DT Flexibility [10-81]	29.78	8.79
Crea.tiles DT Elaboration [4-20]	12.14	2.54
Crea.tiles DT Duration [seconds]	19.58	19.96
Crea.tiles DT Appropriateness [0-4]	1.66	1.20

Elastic net correlations

Tiles CT and DT metrics were combined using an elastic net modeling to predict logic and ideas scores.

Prediction on	Tiles CT	Tiles DT
Logic	$r = .37$, $r_mean = .25$, $p < .005$	$r = .24$, $r_mean = -.008$, $p = .2$
Ideas	$r = .22$, $r_mean = .01$, $p = .8$	$r = .39$, $r_mean = .15$, $p = .02$

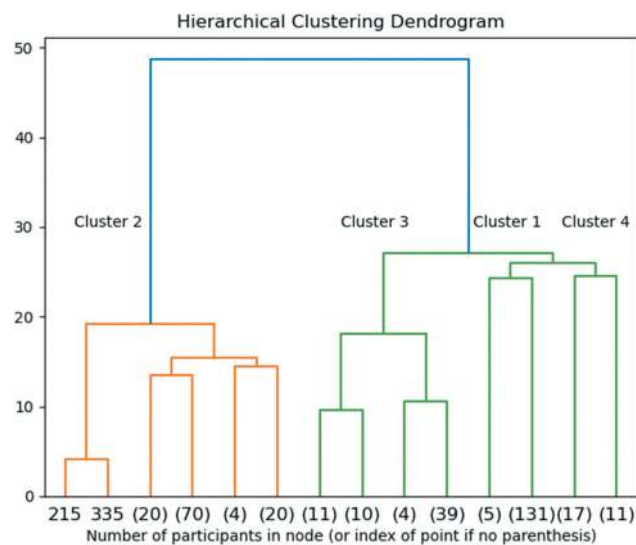
Role of gaming frequency

Correlations between gaming frequency and game metrics.

Player clustering

Dendrogram from AHC algorithm. Four clusters were chosen, annotated below.

Metric	N	r	p
Crea.tiles CT	332	.07	.20
Crea.tiles DT Flexibility	304	-.01	.81
Crea.tiles DT Duration	293	-.04	.46
Crea.tiles DT Frequency	174	-.06	.41
Crea.tiles DT Elaboration	303	.06	.27
Crea.tiles DT Appropriateness	304	.09	.13
Crea.ideas	324	.06	.33
Crea.logic	313	.004	.95



Player clustering visualisation

Cluster visualization via online tool available at <https://vis.scienceathome.org/%20>. As an example, a random

player is shown in cluster 2 (dark green color), who submitted shapes with high appropriateness (dog-like) but low flexibility (small differences between shapes).

