



Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities

Janet Rafner, Michael Mose Biskjær, Blanka Zana, Steven Langsford, Carsten Bergenholtz, Seyedahmad Rahimi, Andrea Carugati, Lior Noy & Jacob Sherson

To cite this article: Janet Rafner, Michael Mose Biskjær, Blanka Zana, Steven Langsford, Carsten Bergenholtz, Seyedahmad Rahimi, Andrea Carugati, Lior Noy & Jacob Sherson (2022) Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities, Creativity Research Journal, 34:1, 28-54, DOI: [10.1080/10400419.2021.1971447](https://doi.org/10.1080/10400419.2021.1971447)

To link to this article: <https://doi.org/10.1080/10400419.2021.1971447>



Published online: 23 Sep 2021.



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



Article views: 953



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 32 View citing articles [↗](#)



Digital Games for Creativity Assessment: Strengths, Weaknesses and Opportunities

Janet Rafner ^a, Michael Mose Biskjær^a, Blanka Zana^a, Steven Langsford^a, Carsten Bergenholtz^a, Seyedahmad Rahimi^b, Andrea Carugati^a, Lior Noy^c, and Jacob Sherson ^a

^aAarhus University; ^bUniversity of Florida; ^cBusiness Administration, Ono Academic College

ABSTRACT

Creativity assessments should be valid, reliable, and scalable to support various stakeholders (e.g., policy-makers, educators, corporations, and the general public) in their decision-making processes. Established initiatives toward scalable creativity assessments have relied on well-studied standardized tests. Although robust in many ways, most of these tests adopt unnatural and unmotivating environments for expression of creativity, mainly observe coarse-grained snippets of the creative process, and rely on subjective, resource-intensive, human-expert evaluations. This article presents a literature review of game-based creativity assessment and discusses how digital games can potentially address the limitations of traditional testing. Based on an original sample of 127 papers, this article contributes an in-depth review of 16 papers on 11 digital creativity assessment games. Despite the relatively small sample, a wide variety of design decisions are covered. Major findings and recommendations include identifying (1) a disconnect between the *potential* of scaling up assessment of creativity with the use of digital games, and the *actual* reach achieved in the examined studies (2) the need for complementary methods such as stealth assessment, algorithmic support and crowdsourcing when designing creativity assessment games, and (3) a need for interdisciplinary dialogs to produce, validate and implement creativity assessment games at scale.

ARTICLE HISTORY

Received April 19, 2021

Introduction

Calls for valid (accurate), and reliable (consistent) methods for assessing creativity can be traced back to Guilford's (1950) presidential address to the members of the American Psychological Association (APA). This call is accentuated in the introduction to the 2019 special issue of *Philosophy, Esthetics, Creativity and the Arts*, "Creativity Assessment: pitfalls, solutions and standards," where the editors assert: "Without proper instruments to measure creativity or adequate standards of assessment, the validity of any creativity study is seriously questioned" (Barbot & Reiter-Palmon, 2019). Recently, organizations such as the OECD (Schleicher, n.d) and the World Economic Forum (Belsky, 2020; Grey, 2016; Future of Jobs Report, 2020) underscored the importance of 21st century skills like creativity for meeting the requirements of an increasingly globalized and digitized world. Calls from these major public institutions triggered an imminent need for *large-scale* assessments and training of these skills, both in corporate and educational sectors. As a student, worker or work-seeker, employer, or governmental agency, the assessment of one's own skill or those of others is paramount for upward mobility and appropriate job placement (Indeed Editorial Team, 2021; Stuart,

2009; Zojceska, 2019). Thus, ensuring the accessibility of assessment mechanisms, such as through freely available online platforms, for creativity and other 21st century skills, becomes crucial for an equitable society. That being said, it can be futile or even harmful to scale-up and provide easy access to invalid or unreliable assessment instruments as is the case with the burgeoning number of apps that promise to evaluate creativity yet lack a solid theoretical foundation.

Over the past decades, standardized tests targeting creative *processes* (e.g., idea generation and refinement) have been developed and successfully deployed in a wide range of contexts (Keefe & Magaro, 1980; Kühn et al., 2014; Lau & Cheung, 2010; Maheshwari & Werd, 2019; Matud, Rodríguez, & Grande, 2007; Torrance, 1976). The simple and domain-agnostic design of the tasks in these tests allows for large-scale, albeit resource intensive, deployment. However, these same design choices are criticized for not being motivational and/or for not mapping well onto real-world creativity (i.e., lacking *ecological validity*¹; discussed later). A parallel line of research has focused on the assessment of creative *products* (Kaufman & Sternberg, 2019). In scenarios exhibiting sufficient relevance and complexity (e.g., music, design, consumer

product) this approach, in conjunction with appropriate expert assessment, has sometimes been hailed as the “gold standard of creativity” (Cseh & Jeffries, 2019, p. 2). Exploring methods that combine the strengths of process-oriented assessments (scalability) and product-oriented assessments (data with high ecological validity) could greatly enhance the construct validity² of creativity assessment instruments.

When compared to traditional instruments, well-designed digital games may be suited to provide this enhanced assessment, as they could allow for the combination of scalability and refined data acquisition. Most research on creativity and digital games focuses on how gameplay fosters creativity (Rahimi & Shute, 2021) and there is relatively little investigation on how creativity assessment can be systematically woven directly and invisibly into the fabric of the game itself, an approach called *stealth assessment* (Green & Kaufman, 2015; Rahimi & Shute, 2021; Shute, 2011; Shute & Rahimi, 2021).

This paper investigates *how digital games can be used to assess creativity at scale, while achieving psychometric validity and reliability*. This investigation is done by means of a systematic literature review, outlining several key design, implementation, and analysis decisions relevant for creating and evaluating digital games which assess creativity. The paper’s primary vision for creativity assessment is to explore ways of meeting the challenge of scaling up to large and diverse sample sizes while capturing data with higher ecological validity and gleaning deeper insights into construct validity of creativity. This paper aims to provide:

- An in depth review of the state of the art of creativity assessment games
- Considerations that can be used to design, implement, and evaluate digital creativity games with respect to validity, reliability, and scalability
- A catalyst for interdisciplinary dialogs

The intended audience for this article is researchers who focus on creativity assessments, as well as stakeholders and practitioners seeking to deploy creativity assessments in their concrete field, such as education, or corporate environments.

Creativity assessment and its challenges: past and present

Creativity is generally defined as the ability to produce solutions, ideas, or products that are both novel and effective (Runco & Jaeger, 2012). Going one layer

deeper, creativity is “the interaction among *aptitude, process, and environment* by which an individual or group produces a *perceptible product* that is both *novel and useful* as defined within a *social context*” (Plucker, Beghetto, & Dow, 2004, p. 91, original emphasis). In the following sections, standard creativity assessment methods will be introduced, and advantages and disadvantages of those will be discussed.

Creative process: convergent and divergent thinking

The majority of psychometric creativity assessment 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) solution to a defined question (Cropley, 2006; Guilford, 1956). Although Guilford drew attention to and acknowledged the importance of both processes, creativity quickly became equated with DT, leading to the majority of assessment instruments focusing on measuring DT (Cropley, 2006). These instruments often operationalize DT following Guilford’s (Guilford, 1973) proposal, in terms of *ideational fluency* (the number of outputs: ideas, products, solutions), *flexibility* (how different the proposed outputs are from each other), *originality* (how unique an output is), and *elaboration* (the level of details in an output).

Standard tests for the creative process

This operationalization of creativity has led to numerous standardized tests, most of which focus on assessing the creative process (Kaufman & Sternberg, 2019, p. 45). What the authors define as ‘creative process’ is equivalent to Guilford’s definition of creative potential (Guilford, 1950, 1968; Runco & Acar, 2012). Guilford’s Alternate Uses Test (Guilford, 1968) asks participants to come up with alternative/unusual uses of everyday objects, such as a brick, or a pencil. A widely used battery (i.e., collection of tests) is the Torrance Test of Creative Thinking (Torrance, 1966), which includes a variety of verbal (e.g., Unusual Uses, Impossibilities, Consequences, Situations) and non-verbal tasks (Incomplete Figures, Creative Design, Circles and Squares). A similar, popular battery is the Wallach-Kogan Creativity Test (Wallach & Kogan, 1965), which includes three verbal, and two visual tasks. Evidence of reliability of these and similar tests is fairly convincing (Kaufman & Sternberg, 2019; Torrance, 1981; Williams, 1980), however the predictive and discriminant (not

correlating with irrelevant measures) validity of DT tests has mixed support (Bachelor, 1989; Clapham, 1996; Cooper, 1991; Kaufman & Sternberg, 2019; Thompson & Andersson, 1983). Some scholars have also noted that scores on divergent production tests are susceptible to training and intervention effects (Clapham, 1996; Kaufman & Sternberg, 2010; Torrance, 1988). CT is commonly measured with the Remote Associations Test (Mednick, 1962), in which a participant is shown three cue words that are linked by a fourth word, which is the correct answer that the participant has to find. While widely used, the Remote Association Test has been criticized for relying heavily on the participant's language skills (Cortes, Weinberger, Daker, & Green, 2019).

Limitations of these standard tests

The aforementioned tests are so well studied that they are often used as a point of reference for convergent validity³ of any newly developed creativity instrument. However, they are criticized for the following reasons:

- They are often administered in strict testing conditions and in artificial contexts (e.g., in laboratories) to ensure internal validity.⁴ These are however far removed from natural situations, where one would express their creativity (Cotter & Silvia, 2019) and can induce unintended physiological effects such as test anxiety (Schram & Ule, 2019) and Hawthorne effect (McCambridge, Witton, & Elbourne, 2014).
- They are not particularly meaningful to the participants (Valladares-Rodríguez, Pérez-Rodríguez, Anido-Rifón, & Fernández-Iglesias, 2016).
- They rely on fairly simple interactions and only analyze the final product, which does not afford the ability to capture detailed, fine grained steps in an individual's trajectory toward an output (Schulte-Mecklenbeck, Kühberger, & Johnson, 2019).
- They rely on resource-intensive administration and human raters (Cseh & Jeffries, 2019). When there are few raters, individual taste, preferences and idiosyncrasies will have an impact. Employing multiple raters and measuring interrater reliability (the extent to which judges agree on the scoring of the answers) is effective but resource-intensive. If raters spend long hours assessing items, they are susceptible to fatigue, compromising the reliability of their judgments (Cseh & Jeffries, 2019).

Additionally, Runco, Abdulla, Paek, Al-Jasim, and Alsuwaidi (2016) compared results from seven different DT batteries and observed that each test seems to measure a slightly different process, and the results are often not highly intercorrelated (i.e., low convergent correlation with each other). This led them to conclude that “research using only one divergent thinking test will not provide generalizable results” (Runco et al., 2016, p. 4). If the purpose of an evaluation is to assess general DT or CT, researchers call for multiple domains and tasks to be included (Acar & Runco, 2019; Cortes et al., 2019; Reiter-Palmon, Forthmann, & Barbot, 2019). However, a significant downside to employing multiple test-like tasks is the increased risk of test fatigue (Ackerman & Kanfer, 2009).

Solving part of the problem: Digital standardized tests

Digital, standardized creativity assessments are increasing in popularity (e.g., Bowden & Jung-Beeman, 2003; Dow & Mayer, 2004; Guo, 2019; Oltețeanu & Zunjani, 2020; Palaniappan, 2012; Pásztor, Molnár, & Csapó, 2015) as they may lower the barrier to algorithmic scoring of responses, thus reducing human subjectivity (Beketayev & Runco, 2016; Lau & Cheung, 2010), and make testing more accessible to administer (Kwon, Goetz, & Zellner, 1998). In principle, the digital nature of tasks in digital tests allows for a more fine-grained data analysis, in which not only the final product, but also the process (e.g., time-ordering of inputs) is taken into account for the evaluation (Evans, Holmes, & Trueblood, 2019). However, this potential is far from being fully exploited in creativity assessment. An example of a promising algorithmic scoring method for assessing DT tasks employs a specific variant of semantic networks, namely an *association network*. Association networks capture the relationships between concepts and words used in natural language (Beketayev & Runco, 2016). Acar and Runco (2014) use a *semantic-based algorithm* (SBA) to calculate how distant (in conceptual space) ideas or answers of an individual are from each other. They found that SBA flexibility is strongly correlated ($r = 0.74$) with traditional scoring of flexibility scores from the Many Uses Test (similar to the Alternative Uses Test; Guilford, 1968), suggesting strong convergent validity. While the approach holds potential, Forthmann, Oyeade, Ojo, Günther, and Holling (2019) highlight that this system is susceptible to certain biases, which can taint the calculations of the system. It must also be noted that automated assessments are far from capturing the complexity and nuance of human judgment needed for some assessments such as scoring originality or flexibility of figural elements of the Torrance

Test of Creativity Thinking. Furthermore, digitized tests are not necessarily more meaningful to the participants than the pen- and paper-based versions.

Product-oriented assessment

In contrast to process-oriented assessments, among others, Baer and Kaufman (Baer, Kaufman, & Gentile, 2004) argue that *product-based* assessments (e.g., poetry, collage, music) are probably “the most appropriate assessments of creativity” as they “indicate ability to produce creative products past or present” (Kaufman & Sternberg, 2019, p. 54). The evaluation of such tasks was operationalized by Amabile in a method known as the Consensual Assessment Technique (Amabile, 1982). This method involves a group of expert raters (i.e., appropriate observers familiar with the domain of the creative product) individually and independently scoring a set of products. The Consensual Assessment Technique has been espoused as a, if not *the* gold standard (Baer et al., 2004) in creativity assessment and considered a versatile, reliable measure of creativity negating the need for explicit definitions (Cseh & Jeffries, 2019). This technique relies completely on resource-intensive human rating, especially as expert raters are sometimes hard to find. However, the requirements for one to be an expert, and information on training or guidance of judges is not well defined (Cseh & Jeffries, 2019). This uncertainty, coupled with rapid technological developments, has led to initial studies of crowdsourced consensual assessment, however the results have been mixed (see McKenna, Arnone, Kaarst-Brown, McKnight, & Chauncey, 2013).

Other traditional testing instruments

Before moving on to discuss digital game-based assessments, one should note that there are also well-known assessments for creative personality, behavior, and environments. Aspects of personality and behavioral correlates of creativity are evaluated with self-perception assessments such as personality scales and questionnaires (Kaufman & Sternberg, 2019). Self-reporting can be useful in some contexts despite its vulnerability to biases (Kaufman, 2019). Creativity can be assessed from observing a particular activity, which is common practice in education (Katz-Buonincontro & Anderson, 2018). One can map the process toward a creative solution by recording verbal transcripts of thought. One commonly used method employing this is the “Think Aloud” method (Dorst & Cross, 2001; Kaplan & Simon, 1990). However, these are also resource intensive to manually code.

Digital game-based assessment of creativity

In contrast to traditional instruments, well-designed digital games may be suited better for assessing creativity for many reasons. Broadly, game-based assessment can be described as “using data from in-game actions to make inferences about players’ knowledge, skills, and attributes, allowing us to make use of information in the ocean of data produced by daily digital interactions with software” (DiCerbo, 2014, p. 26). In the context of creativity assessment, there are several factors which could allow for the combination of scalability and refined data acquisition. First, the gamified environment elicits behavior that is more casual and natural than the completion of traditional tests. This can mitigate unintended physiological effects, such as test anxiety, or researcher effect and in turn may result in data with higher ecological validity (Lumsden, Edwards, Lawrence, Coyle, & Munafò, 2016; Shute & Kim, 2014). Digital games are also more motivating to people than standardized tests are (Gee, 2005). Intrinsic motivation can induce a ‘flow state,’ in which people fully concentrate on and enjoy the task-at-hand. This state can be achieved when people have clear goals, immediate feedback, and a balance between challenges and skills (Csikszentmihalyi, 1997). Being in the flow-state also allows for creativity (Csikszentmihalyi, 1997), and thus creates a good opportunity for creativity assessment.

Additionally, digital games enable higher engagement levels. This could allow for longer interactions and multifaceted testing of creativity in a complex environment, which is necessary to elicit and assess both creative processes and products, without inducing testing fatigue. This point is paramount in creativity research, as more and more emphasis is placed on the inclusion of multiple domains and tasks, in order to adequately make generalizations about an individual’s creativity. Digital games also afford the ability to capture detailed, fine grained steps in an individual’s trajectory toward an output (Hart et al., 2017). Given their accessibility, appeal and prevalence in peoples’ everyday lives (Ghuman & Griffiths, 2012) digital games could allow for broader sampling (engaging many demographic groups), possibly leading to *reliable* predictions at the individual level (through validation of demographically fine-grained models), higher external validity, and more generalizable results. On the one hand, if the games are freely accessible online, there are several risk factors from lack of research supervision such as people playing without paying attention. On the other hand, freely accessible digital games could support the equitable access to creativity assessment. These factors underscore the potential digital games hold for assessing creativity.

In the absence of valid, reliable, scalable, and accessible solutions for assessing creativity, non-scientifically driven corporate enterprises have leveraged new technologies, creating widely used games and gamified platforms. These games claim to foster and/or assess creativity and other 21st century skills. Examples include: *KnackApp* (u.d.), *Kettlemind* (u.d.), *Figment* (u.d.), *BrainSparker* (u.d.), and *Lumosity* (u.d.). Some of these companies charge fees based on promises to “find hidden talent and build your strongest workforce” (*KnackApp*, u.d.), and encourage people to purchase and play so they “do not get left behind” (*KnackApp*, u.d.). These games provide a user friendly ‘creativity score’ which allows one to ‘certify’ their creativity. Very few of these describe transparent methodologies or any scientific documentation for their claims, and the validity of some that do have been seriously questioned (Shute, Ventura, & Ke, 2015). Despite their caveats, these apps maintain substantial user bases in the tens of thousands to millions (Strobach & Kupferberg, 2020). This information truly highlights the imminent need for accessible, scalable, scientific, and transparently documented solutions.

Fortunately, efforts to systematically analyze and assess elements of creativity and “processes that can evolve into creative solutions” (Shute & Rahimi, 2021), embedded in technology-rich environments (e.g., digital games) are underway. Assessment in these environments is called *stealth assessment*, where the player’s in-game actions are continually logged, while the participant is immersed in the activity, without being constantly attune to the assessment. Stealth assessment is rooted in *evidence centered design* (Mislevy, Almond, & Lukas, 2003), which combines knowledge from both computational, and conceptual models. This method requires the researcher to “(a) define the claims concerning learners’ competencies, (b) establish what represents valid evidence of a claim, and (c) determine the kind of tasks or situations that will elicit that evidence” (Shute & Kim, 2014, p. 107). Stealth assessment systems use two main processes, which incorporate the requisites of evidence centered design. These are evidence identification (i.e., identifying the predefined variables or observables from log files), and evidence accumulation (i.e., statistically linking the observables to the facets of the competency at hand), which allow for the assessment of the player’s competency level in real-time. Assuming that the underlying theory is robust, the more data the players generate, the more accurate the estimate for each player becomes. Additionally, the fine-grained tracking and recording of player’s interactions (i.e., via the accumulation of log data generated by keystrokes, mouse clicks, and choice patterns) can be used to analyze the player’s learning progress and provide ongoing feedback (Shute, Lu, & Rahimi, 2021).

Given the combination of rich, fine-grained data and versatile task design, stealth assessment has been applied to target “hard-to-measure constructs,” including creativity (Shute & Rahimi, 2021), basic and executive cognitive functions (Pedersen et al., 2020), problem solving (Akram et al., 2018; Shute et al., 2015; Shute, Wang, Greiff, Zhao, & Moore, 2016), persistence (Rahimi, Shute, & Zhang, 2021), computational thinking (Min et al., 2015), social skills (DeRosier, Craig, & Sanchez, 2012), and conscientiousness (Shute & Ventura, 2013). In general, any technology-rich environment that includes human-computer interaction and requires the assessment of a particular knowledge, skill, or attribute should be amenable to the application of stealth assessment. For a comprehensive review on stealth assessment see (Shute, Lu, Rahimi, 2021) and (Shute & Moore, 2017).

Summary of assessment methods

Main principles, pros and cons of assessment methods discussed before, are summarized in the Table 1. As it appears from the table and the argumentation above, there is no perfect way to measure creativity at scale. The use of digital games is one of the latest innovations to measure creativity, and while they seem promising, their novelty warrants some level of scrutiny. The following literature review focuses on digital game-based mechanisms such as stealth assessment, addressing the research question: *how can digital games be used to assess creativity at scale, while achieving psychometric validity and reliability?*

Methods

A literature review and iterative coding exercise was performed in order to understand how digital games measure creativity. The following section describes the review’s sampling methodology and coding scheme.

Sampling

In this review, papers that report on digital game-based creativity assessments were included regardless of their scoring mechanism (human or algorithmic). While only the three papers written by Shute et al. use the terminology ‘stealth assessment’ and ‘evidence centered design’ explicitly, all papers build on a theoretical foundation of creativity (mostly divergent thinking or convergent thinking), operationalize it given their experimental set up (e.g., explain their metrics for flexibility, fluency and originality), and collect data from the game play, based on the operationalization. Papers that described *physical* (as opposed

Table 1. Summarizes the main principles and pros and cons of the creativity assessment instruments discussed in this paper

Assessment method	Main principles	Pros	Cons
Pen and paper standardized process oriented assessments	Assesses creative processes through individual or batteries of tests targeting components of DT or CT	- Large corpus of research - Decades of practical use - Often easy to manually administer - Supervised by the researcher, ensuring focus	- Resource intensive scoring via human raters and thus cumbersome to scale - Potentially unmotivating - Ultimately captures only final products and not the process - Simplistic task design affords primarily coarse grained data - May induce unintended psychological effects
Digital standardized process oriented assessments	Assesses creative process through individual or batteries of tasks targeting components of DT or CT	- Scalable if automated assessment is possible	- New and not widely studied yet - Potentially unmotivating - Resource intensive scoring when human raters are employed, and thus cumbersome to scale - May induce unintended psychological effects
Product oriented assessment	Assesses the creativity of a product by individual and independent consensus of expert raters	- Large corpus of research - Decades of practical use - Clearly indicates ability to produce creative products past or present	- Resource intensive and therefore often not easily scalable
Person and environmental focused assessment	Includes assessment of creative behavior, creative self perception	- Large corpus of research - Decades of practical use	- Increased risk of confounding variables given the broad scope - Bias due to self reporting - Increased risk of - Resource intensive and therefore often not easily scalable
Commercial game based assessments	Can in principle be forms of digitized psychometric and non psychometric tests,, or other unspecified mechanisms	- Designed to be used at scale - Engaging interaction design	- Little (if any) scientific backing or disclosure of methodology - Costs limit equitable access
Digital game employing stealth assessment	Evidence needed to assess the creative process is provided by the players' interactions with the game itself	- Could offer opportunities for higher ecological and construct validity - Scalable when automated assessment is used - Potentially more motivating than standard tests - Afford fine grain metrics	- New and not widely studied yet

to digital) games (e.g., board games, role-playing games, sports) were excluded, alongside studies writing about how digital games *foster* (as opposed to measure) creativity, and articles that focused on investigating the creative process of *designing* (as opposed to playing) video games.

Multiple sources were used in this review. Given the historical roots of modern creativity research (i.e., since Guilford's (1950) inaugural presidential APA address), APA-based outlets were first searched. However, as Gardner wrote, creativity is "precisely the kind of problem which eludes explanation within one discipline" (Gardner, 1988, p. 22), the scope of sources was expanded to include other affiliated disciplines and their main publication channels. The final list of libraries and journals included: Creativity Research Journal, PubMed, ACM, Wiley, the APA library, and Google Scholar. All libraries were searched using the following queries in the abstracts of the papers: '*creativity AND games AND assessment*' as well as '*creativity*

AND games AND measurement.' The only exception was Google Scholar, which does not have a function for searching in abstracts. Sources, excluding Google Scholar, returned 103 publications fitting the search terms. Since Google scholar returned 200,000+ publications and in order to make this task comprehensible, titles and abstracts for pages 1–10 were skimmed through (each page listing about 10 hits), as after around page 6 there were no relevant hits. This was done for both search terms, resulting in 193 hits in total. 109 paper were excluded based on irrelevant titles (e.g. "The effects of non-specific and specific concepts on tactical creativity in team ball sports"). 60 publications were excluded based on reading the abstracts. 24 publications were duplicates of papers already found in other libraries. Eight papers that were found prior to the review and fit the criteria but did not appear in the search were manually added. Sampling took place between November 2020 and January 2021.

Analysis

The process described above resulted in a pool of 127 studies in total. These papers were manually read and evaluated by the first and third authors to determine if they fit the above stated criteria. Papers that did not fit the criteria, and duplicate papers were removed from the review. In total, 16 papers fit the review's criteria (digital games used to assess creativity, regardless of their scoring mechanism). Google Sheets were used to organize and collaboratively discuss the resultant publications among the coauthors.

The coding scheme was developed through an iterative process that involved discussions among all coauthors. As a first naive approach, open-ended data was collected on *research focus* and *definition of creativity* in order to get a broad sense of the studies. Then granularity was increased to include the *game description*, *scoring methods* used, the number of *participants*, whether individual, *dyadic*, or *group* creativity was being assessed, the *stated motivations* for using game-based assessments, the method of *validating* the game, and the level of *correlation* observed. Some seemingly straightforward codes such as the *number of participants* proved to be problematic. For example, due to anonymized data, one study reported the number of game sessions, not the number of players. With regards to the focus of the study being individual, dyadic or group creativity, an additional option of *co-creative* was added, when a paper describing a game that a child plays together with a robot was encountered. Following a variation of Remy, MacDonald Vermeulen, Frich, Biskjaer, and Dalsgaard (2020), codes were systematically divided into three categories: open ended, numeric, and nominal.

The next iteration of codes was inspired by fundamental questions one might ask while designing a digital game to assess creativity. For example, in addition to the definition of creativity, the component of creativity measured (e.g., DT-flexibility, DT-fluency, DT-originality, CT) was added. In addition to the *number of participants*, the codes of *location*, *expertise*, *recruitment method* and *compensation* was added, in order to address the issues of scalability and reaching a diverse pool of participants. The publication outlets were also recorded and sorted to glean insight into the diversity of domains exploring game-based creativity assessment. The open code *venue of publication* was added to highlight the diversity of publication outlets. See [table 2](#) for the final list of codes used in the analysis.

Motivation for using games

Each paper was searched for explanations as to why games had been used to assess creativity and the reasons authors provided were identified. The papers

were blindly coded by the first and third author and then through iterative discussions, the motivations presented in the papers were coalesced into seven categories. Data richness, player motivation and engagement, and scalable assessment were the three most commonly cited reasons for using digital games.

Game design: mechanics and metrics

The types of games used in the reported studies varied widely. Three studies designed their experiment around preexisting, commercially developed games (Chuang et al., 2015; Gal et al., 2017; Voiskounsky et al., 2017), while all other studies used games that had been designed for academic purposes. The core game play interactions varied from sandbox games such as Minecraft (Voiskounsky et al., 2017) through image blending with neural networks (Rafner et al., 2020), to educational games focused on Newtonian physics (Shute & Rahimi, 2021).

The games also differed in their inclusion of various levels of difficulty, progression, multiple rounds, or multiple modes of play. Here, a 'mode of play' refers to different types of tasks within a game, while 'rounds' refer to the number of times a participant completes the same type of task (no varying difficulty) and levels refers to similar tasks with increasing difficulty. For example, *crea.blender* (Rafner et al., 2020) includes three modes, Create, Challenge and Open Play, all of which are different types of tasks with same core game play, image blending with a neural network. In contrast, the reviewed study on Minecraft (Voiskounsky et al., 2017) consists of two rounds of a single mode, with participants instructed to build two objects, a house and a ship, which can be considered to require the same competencies. Overall, four games implemented progression explicitly in their design, out of which, three worked with level-based, and one with time-based progression.

Two of the games explore digitally augmented creativity by allowing players to interact directly with an algorithm in the process of game-play. In *crea.blender*, participants interact with a neural network to create images, and in *Droodle* (Ali et al., 2019), the participants interact with a social robot by taking turns playing the game.

As seen in [Table 5](#), seven of the games measure DT through some combination of fluency, flexibility, and originality, while only *crea.blender* explicitly measures CT (Rafner et al., 2020) but does not provide metrics for those measures. The Creativity Game-based Evaluation System (Yeh et al., 2016) utilized insight problems which could be solved with different paths. While the process of solving insight problems is complex, and requires both CT and DT (DeYoung, Flanders, & Peterson, 2008), the

study only reports and measures creativity through a binary variable, the correctness of the solution. BuzzWords (Howell & Veale, 2009) reports on verbal creativity and word association which is commonly linked to CT. In Immune Defense (Krebs et al., 2020), creativity is calculated through the measurement of diversity in the player's solutions toward the same goal.

Results and synthesis

In this section, the results from the coding exercise are described, and the prescriptive findings are presented. As a point of clarification, from the 16 papers included in the review, 11 unique games (Ali, Moroso, & Breazeal, 2019; Chuang, Liu, & Shiu, 2015; Gal, HersHKovitz, Morán, Guenaga, & Garaizar, 2017; Hart et al., 2017; Howell et al., 2010; Krebs et al., 2020; Rafner et al., 2020; Shillo, Hoernle, & Gal, 2019; Shute & Rahimi, 2021; Voiskounsky, Yermolova, Yagolkovskiy, & Khromova, 2017; Yeh, Lai, & Lin, 2016) were identified. An overview of the results from the nominal and numeric codes can be seen in Appendix B, Table B2. Screen shots of each game can be found in Appendix A, Figures A1-A10

Scientific context: research questions and publication outlets

All studies were coded on their research questions and approach. The research approach was either empirical, when data collection was involved, or theoretical, when there was no data involved. In total, there were 13 empirical, and 3 theoretical papers. Among the empirical studies, it was possible to distinguish between usability studies (participant's user behavior and interaction with the tool was observed; Rafner et al., 2020), experimental (creativity was studied under two or more experimental conditions; Ali et al., 2019; Noy et al., 2012; Voiskounsky et al., 2017; Yeh et al., 2016), correlational (investigated which variables have a natural relationship with creativity; Gal et al., 2017; Rozenkrantz et al., 2017; Voiskounsky et al., 2017; Yeh et al., 2016), and validation (tested for correlation with preexisting, standard measures; Chuang et al., 2015; Hart et al., 2017; Howell & Veale, 2010; Krebs et al., 2020; Rozenkrantz et al., 2017; Shute & Rahimi, 2021) studies (some belonging to multiple categories). Among the theoretical papers, it was possible to differentiate between descriptive (described a tool/proposal; Howell & Veale, 2009; Shute & Wang, 2016), and comparative studies (analyzed/compared ideas/methods; Almond, Kim, Velasquez, & Shute, 2014).

To further document the diversity of studies, publication outlets were noted. Out of the 16 papers, eight were published in eight different conferences, seven were published in five different journals, and one was a book chapter. These results can be found in Tables 3,4.

The papers were searched for the elements that made the authors characterize the activity as a *game* instead of simply a digital task. All papers described their participants as "players" and all except for Open Creativity App (Shillo et al., 2019) explicitly referred to their project as a "game." Some games used narratives (i.e., a context description beyond what is necessary for the execution of the task) to induce a game-like environment. These included three projects that used a preexisting game, such as Kodetu (Gal et al., 2017), in which players had to "guide the astronaut to the end of a maze," Minecraft (Voiskounsky et al., 2017), and Arctic Quest (Chuang et al., 2015), and one self-designed initiative, the Creativity Game-based Evaluation System (Yeh et al., 2016), which used the context of finding a way out from a kitchen or a living room. BuzzWords (Howell & Veale, 2009) and Immune Defense (Krebs et al., 2020) were designed to resemble existing games. BuzzWords is a Tetris-like word game, while Immune Defense is designed as a tower defense game. Other projects used the game instructions to create an intrinsically motivated playful environment. For instance, in the Droodle Game (Ali et al., 2019), a social robot asked kids if they wanted to play, while in crea.blender (Rafner et al., 2020), Physics Playground (Shute & Rahimi, 2021), Creative Foraging Game (Hart et al., 2017), and the

Table 2. Final codes used in our analysis.

Type	Codes ^a
O	Motivation
O	Research question(s)
O	Game description
M	Individual
M	Dyads
M	Group
M	Co-creative
O	Component of creativity measured
O	Method of assessment
O	Validation mechanism
N	Convergent correlation
N	Duration in hours
N	Number of participants
O	Level of expertise of participants
O	Compensation for participants
M	Location: lab
M	Location: school or university
M	Location: remote (online)
M	Location: unspecified

^a0 = open ended; N = numeric; M = nominal.

Table 3. Publication outlets for included studies.

Conference Title	Ref.	Journal Title	Ref.	Book Title	Ref.
Creativity and Cognition	Ali et al., 2019	Computers in Human Behavior	Shute et al., 2021; Yeh et al., 2016	The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications (Chapter 21)	Shute & Wang, 2016
Computational Creativity	Noy et al., 2012	Multimedia Tools and Applications	Chuang et al., 2015		
Foundations of Digital Games	Howell & Veale, 2009	PLOS One	Hart et al., 2017; Rozenkrantz et al., 2017		
Learning At Scale	Gal et al., 2017	Psychology in Russia	Voiskounsky et al., 2017		
Educational Data Mining	Shillo et al., 2019	Measurement: Interdisciplinary Research and Perspectives	Almond et al., 2014		
Engaging Pedagogy	Howell & Veale, 2009				
Computer-Human Interaction in Play	Rafner et al., 2020				
IEEE Conference on Games	Krebs et al., 2020				

Table 4. Codes and their descriptions for motivations for using games for creativity assessment.

Motivation	Description	Reference
Scalable assessment	Given by calls for automated evaluation of creativity in contrast to manual assessment	Shute & Rahimi, 2021; Gal et al., 2017; Rozenkrantz et al., 2017; Shillo et al., 2019; Krebs et al., 2020; Hart et al., 2017
Increased sample size	The study can reach a wider, deeper population sample	Gal et al., 2017; Krebs et al., 2020; Shute & Wang, 2016
Objective Assessment	Automatized assessment eliminates certain levels of subjectivity in creativity assessment.	Chuang et al., 2015; Krebs et al., 2020; Shute & Rahimi, 2021
Ecological validity of data	Data collected from a game environment may provide data that reflects real-life behavior more authentically by, for instance, reducing test anxiety and researcher bias	Rafner et al., 2020; Voiskounsky et al., 2017; Shute & Wang, 2016; Krebs et al., 2020
Data richness	Games allow for the tracking of progressing and more refined and richer data	Howell & Veale, 2009; Rafner et al., 2020; Shute & Rahimi, 2021; Ali et al., 2019; Almond et al., 2014; Shute & Wang, 2016; Rozenkrantz et al., 2017; Voiskounsky et al., 2017; Krebs et al., 2020; Hart et al., 2017
Player motivation and engagement	Games can have a positive effect on the players themselves such as mention that games can reduce a player's test anxiety be more motivating and enjoyable than traditional tests	Rafner et al., 2020; Howell & Veale, 2010; Howell & Veale, 2009; Almond et al., 2014; Shute & Rahimi, 2021; Yeh et al., 2016; Hart et al., 2017; Ali et al., 2019
Familiarity with games	Given by citations that games are common in everyday life, thus they are a convenient mode of data collection for the participants.	Shute & Wang, 2016; Shute & Rahimi, 2021; Chuang et al., 2015

Open Creativity App (Shillo et al., 2019), instructions encouraged players to “explore” and “discover,” various “beautiful,” “awesome,” and “creative” solutions. Two studies collected information on the subjective experience of “playfulness” (Rafner et al., 2020) and “enjoyment” (Shute & Rahimi, 2021). See Table B1 in appendix B for the full instructions given to the participants for each of the studies. See appendix A for screenshots of the games from the reviewed papers.

Scoring mechanisms and instrument validation

The sampled studies apply a diverse range of analyses appropriate to their wide variety of goals and subject matter domains. A primary distinction was made between analysis methods that apply manual human rating and those that employ automated scoring of in-

game creativity. Among human rating methods, it is possible to distinguish between self, peer (other players), and expert judgment, as is traditionally done in creativity research (Sawyer, 2011). Among the games, Doodle (Ali et al., 2019) and Minecraft (Voiskounsky et al., 2017) used human rating only in the form of expert raters (although Doodle calculated the novelty using automated Natural Language Processing methods). The rest of the games calculated creativity automatically, using various in-game information and calculation approaches such as Bayesian Networks (Shute et al., 2021). BuzzWords (Howell & Veale, 2009; Howell & Veale, 2009) categorized solutions as low- medium- or high creativity based on compound word-frequency, informed by WordNet (Fellbaum, 2010). Others employed automated scoring based on some arbitrary baselines of creative behavior (e.g., expert input, previous gameplay data, computer-

Table 5. Short description of games, their context and type of creativity measured.

Game	Short description	Commercial or academic	Assessed components of creativity
Kodetu variant ⁵ (Gal et al., 2017)	Each of Kodetu's 15 levels presents the user with a maze in which an astronaut should get to a marked destination; guiding the astronaut is done via a block-based code the user is editing.	Commercial	Fluency (number of solution attempts), Flexibility (strategy used), Originality (based on the distribution of solutions)
Arctic Quest 2 variant (Chuang et al., 2015)	In Arctic Quest 2 players must take falling pieces and place them correctly on the marks according to their shapes. Players earn trophies and medals for completing challenges in the game.	Commercial	Fluency, Flexibility, Elaboration, Sensibility
Minecraft variant (Voiskounsky et al., 2017)	Minecraft is a commercial sandbox video game developed by Mojang. Players explore a blocky 3-D world and discover and create items and structure	Commercial	Productivity (i.e., the number of ideas, goals created), Originality
crea.blender (Rafner et al., 2020)	In crea.blender, players interact with a neural network to blend together source images into new creative images. Game contains 4 modes: a tutorial, a DT mode, CT mode and open play mode.	Academic	CT, DT
BuzzWords (Howell & Veale, 2009, 2010)	The game is a word-stacking exercise in which the player decides how words are best combined: word-combinations that form known compound terms (e.g., "roast" + "beef") obliterate each other to yield a bonus score, while those that do not combine stack up to clog the game screen. There can be several matches, but only one will be the best or rarest match. Over time the difficulty (or rarity) of the words increases or decreases based on the player's performance.	Academic	Verbal creativity/word- association skills (number of moves, total score, maximum difficulty level achieved and average difficulty level throughout the game)
Physics Playground (formerly Newton's Playground) (Shute & Rahimi, 2021; Shute & Wang, 2016; Almond et al., 2014)	Physics Playground is a computer-based game designed to assess and support students' conceptual understanding of physics principles. In <i>PP</i> , players draw various objects on the screen using a mouse or stylus, and once drawn, these objects become "alive" and interact with other objects.	Academic	DT: flexibility, fluency, and originality
Other: relevance, originality, esthetics, humour/ surprise Doodle (Ali et al., 2019)	Participants are presented with a Doodle, a simple abstract drawing that "comes into focus" (in a surprising way) with the addition of a clever title, and they are tasked with generating Doodle titles. The child and the robot take turns to play the game.	Academic	Fluency, Novelty, Value
Creative Foraging Game (Hart et al., 2017; Rozenkrantz et al., 2017; Noy et al., 2012)	Participants move ten adjacent squares to create different shapes. The players are instructed to save shapes that they find interesting and beautiful into a gallery.	Academic	DT: Fluency, flexibility, originality
Open Creativity App (Shillo et al., 2019)	Participants are requested to create as many different polygons types as possible by intersecting two rectangles.	Academic	DT: Fluency, Flexibility, Originality
Creativity Game-based Evaluation System (Yeh et al., 2016)	The CGES included a series of insight problem tasks that were adapted from the Situation-based Creativity Tasks (SCT) to evaluate the participants' game-based creativity in this study.	Academic	Correctness
Immune Defense (Krebs et al., 2020)	In Immune Defense, participants can place and upgrade objects and can combine object effects, in order to protect a "life form" from enemies. The game is implemented as a tower defense game, a common sub-genre of strategy video games, using an open-source game engine.	Academic	Diversity

Note: three papers from the reviews studied Physics Playground, three papers studied the Creative Foraging Game, and two papers studied BuzzWords.

generated solutions) (Ali et al., 2019; Chuang et al., 2015; Gal et al., 2017; Hart et al., 2017; Krebs et al., 2020; Noy et al., 2012; Rozenkrantz et al., 2017; Shillo et al., 2019; Shute & Rahimi, 2021; Voiskounsky et al., 2017; Yeh et al., 2016). The methods for automated scoring of responses varied so dramatically due to the diversity of the game designs, that it was not meaningful to cluster them, but they will be mentioned in the discussion.

Six out of 11 games tested the convergent validity of their instruments by testing their correlation with traditional measures. These results are summarized in Table 6.

Evaluation details: Duration, location, participants

Out of 16 papers, 12 included information on the duration of the gameplay and evaluation sessions. Chuang et al. (2015) reported that participation spanned two weeks without giving the amount of time spent on gameplay. For all other reporting papers, interventions lasted between eight minutes and four hours (average 50 minutes). Only six studies noted information on the location of the study explicitly. These varied between school (2), research laboratory (2), and online platform (2). Twelve papers reported the number of participants, ranging from eight to 266, with a mean of 85 (SD = 79.4). The participant pools were students from varying grades of elementary school (four studies), university (six studies), or general population (one study). One study used MTurk to recruit participants, two studies mentioned reaching participants through e-mails or social media, and one study recruited people on a snow-ball basis. Six studies reported some form of compensation (money, gift card, sweets) in exchange for participation.

Discussion

The following sections discuss the results and synthesis of the review as well as provide recommendations for future work.

Size matters and so does accessibility

From the analysis, a clear gap can be identified between the *potential* of scaling up assessment of creativity with the use of digital games, and the *actual* reach achieved in the examined studies. It is important to differentiate between two distinct purposes of scaling up. In research driven studies, scale up typically involves crowdsourcing and public participation through efforts to enhance engagement; the benefits of increased number of participants may include the possibility of posing and validating increasingly granular scientific models or group comparisons through the increased statistical power (Meehl, 1967). Thus, a given model granularity for e.g. an assessment tool sets a threshold requirement for participant numbers to achieve validity. After that is achieved the tool is ready for *implementation scale-up* that only aims to broaden the accessibility of the tool, not to provide further research validity. Such scale-up will typically be pursued through the established primarily educational networks such as inclusion in international or national educational testing initiatives.

Out of the reviewed papers, the highest number of participants was 226 (Yeh et al., 2016) with an average of 85 per study (See Figure 1). The reasons for the relatively low sample sizes in these studies could be due to the fact that achieving large sample size was not a priority, that the presented games are still in the early, laboratory stages, or due to the challenges of scaling up (e.g. technical and recruitment challenges). To elaborate on this, only three studies reported that scaling up is one of the primary reasons for using games (Shute & Wang, 2016; Gal et al., 2017; Krebs et al., 2020). Moreover, the nature of studies can determine the possibility for scaling up. For example, in Ali et al. (2019), participants physically interact with social robots which makes scaling up much more difficult than in the purely digital studies. Furthermore, Krebs et al. (2020) states in the abstract that the work is “preliminary data” and then in the conclusion emphasizes several times the need for further studies with more participants.

Table 6. For each game reviewed, Sample Size (N), Standardized assessment instrument (Test), Correlation Coefficient (r) between the standardized assessment instrument and the game. Stars indicate significance, * $p < .05$, ** $p < .01$. RAT refers to the Remote Associates Test, W&K to the Wallach Kogan Test, Williams CAP to the Williams Creativity Assessment Packet, and AUT to the Alternate Uses Test.

Game	N	Test	Correlation	Reference
Buzzwords	26	RAT	$r = 0.65^{**}$	Howell & Veale, 2009
Physics Playground	167	W&K Fluency	$r = 0.18^*$	Shute & Rahimi, 2021
		W&K Originality	$r = 0.18^*$	
Arctic Quest 2	30	Williams CAP	$r = 0.81^{**}$	Chuang et al., 2015
Creative Foraging Game	57	AUT Fluency	$r = 0.18$	Hart et al., 2017; Rozenkrantz et al., 2017
		AUT Originality	$r = 0.20$	
Immune Defense	17	AUT Fluency	$r = 0.54^*$	Krebs et al., 2020
		AUT Flexibility	$r = 0.57^*$	
		AUT Originality	$r = 0.13$	

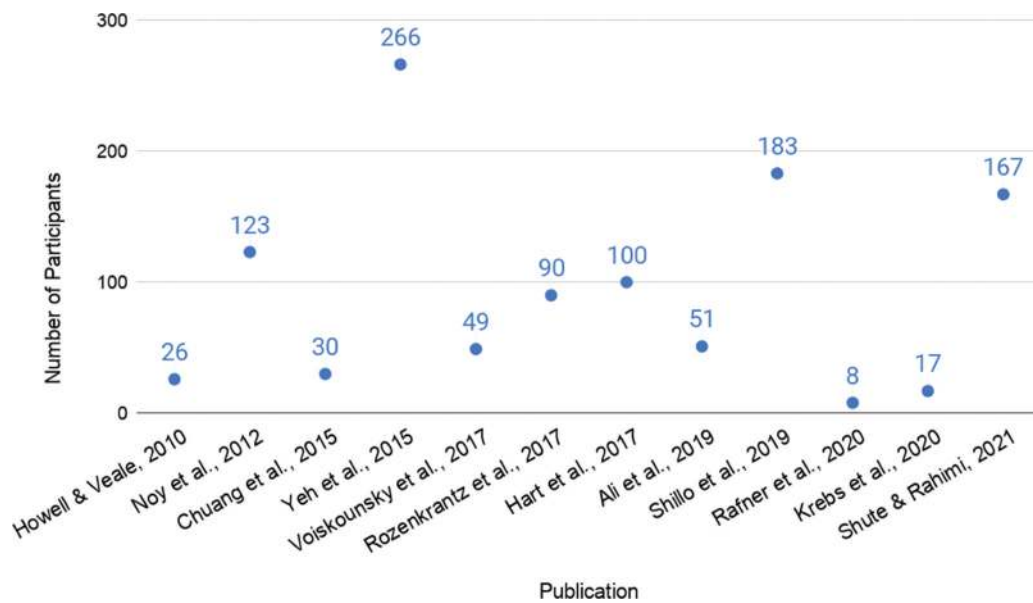


Figure 1. Participant Numbers in 12 publications (one study did not disclose number of participants, three papers refer to a single study of Physics Playground and are thus represented only once, two papers discuss BuzzWords but only one of them refers to an experiment).

With regards to the challenges of scaling up, evidence is mounting from previous and ongoing crowdsourcing games (e.g., Sea Hero Quest, Foldit, Skill Lab) that they can, in principle, elicit massive recruitment, but only a small fraction of projects achieve this “viral” status (Hyde et al., 2016; Koepnick et al., 2019; Pedersen et al., 2020). For this to happen, a number of factors have to be fulfilled simultaneously, including high fun factor, clear sense of purpose, engaging interaction design, sufficiently compelling graphics and narrative, and substantial marketing campaign.

Traditional behavioral studies are conducted in labs or other relatively controlled environments to maintain *internal validity*. In the lab, the participant is supervised to ensure comprehension of the instructions and attention to the activity (Casaletto & Heaton, 2017). These types of studies find participants through open calls (flyers, etc.), via teaching, and often offer compensation. In many cases social scientists have turned to paid, recruiting platforms such as mTurk and Prolific in order to scale up sample sizes (Casler, Bickel, Hackett, 2013); however, these platforms also have several drawbacks (Aguinis, Villamor, & Ramani, 2021; Huber & Oded, 2017). For example (Pedersen et al., 2020), compared data from mTurk participants and data from volunteer players in a series of game-based cognitive ability tasks, showing detrimental trends in parts of the mTurk data set where participants sacrificed accuracy

for speedy task completion. For creativity tasks in particular, the collection of data via paid platforms may be particularly challenging because, compared to many basic psychology tasks, they typically exhibit a higher degree of open-endedness, which can make it harder to create alignment between scientifically productive and pay-off maximizing solution strategies (Dontcheva, Gerber, & Lewis, 2011; Kittur, 2010; Oppenlaender, Milland, Visuri, Ipeirotis, & Hosio, 2020). Volunteer-based outside-of-the-lab studies, in which participants contribute at their own leisure (e.g., using their phone or computer), allow for millions of participants, but at the cost of losing some of the control over the testing environment. Fully understanding and finding ways of overcoming this loss of control represents one of the most urgent challenges of volunteer-based outside-of-the-lab assessment. For these studies, crowdsourcing and gamification can be used to increase the bases of contributors and expand areas of scientific inquiry (Beck et al., 2020). However, the game(s) must still overcome the aforementioned challenges to attract and sustain volunteer participants.

As to the accessibility of the games, additional research showed two (Hart et al., 2017; Shute & Rahimi, 2021) out of the eleven games are, at the time of this study, available to play online. However, neither of these games present the player with feedback on their creative abilities based on their game-play. For certain behavioral studies, it can be necessary to keep the

instrument offline, out of sight of the public, so as to not pollute the experimental data when assessing a particular intervention. However, if one is to attain a truly large scale and realize the benefits of normative analyses for heterogeneous populations, online availability seems to be paramount. Online availability is also important so that multiple stakeholders (e.g., governments, organizations, individuals) can reap the benefits of widely accessible creativity assessments. As mentioned above, the lack of the scientifically founded publicly available creativity assessments has given rise to a somewhat wild west of commercial offerings similar to that of brain training games and apps.

Given the call for growing amounts of increasingly complex player data and difficulty in creating ‘viral’ games, one may ask whether the existing immense sources of commercial games might present a promising path forward. To illustrate that this is feasible, Shute et al. (2016) described their process of implementing stealth assessment of problem-solving skills in the popular game, *Plants vs. Zombies 2*.

While another promising line of research, the methodological challenges of pursuing such an approach remain daunting: (1) either the gameplay data is not freely available or the appropriate data is not recorded in the log files, (2) in order to achieve alignment with existing creativity research either games with suitable subtasks have to be identified or the game developers convinced to insert these modes into the game, and finally, (3) in order to enable reliable testing conditions in outside-of-the-lab studies, one would ideally structure conditions so that many players go through roughly the same dynamics, which is not necessarily in agreement with commercial interests in maximizing player freedom of choice and enjoyment.

Game complexity and its relation to ecological- and construct validity

Next, game elements which support higher ecological- and construct validity are considered. As discussed above, well-designed digital games can facilitate the ‘flow state’ when there are clear goals, immediate feedback and an appropriate balance between challenges and skills (Csikszentmihalyi, 1997). The flow state is important for creating an environment which affords the creativity of the players so creativity can be assessed. Additionally, just as researchers assert that multiple tasks are needed to assess elements of creativity (Acar & Runco, 2019; Cortes et al., 2019; Runco et al., 2016), it can be asserted that progression through game levels, modes and rounds are also needed in order for digital games to supply sufficiently comprehensive and diverse data to accurately and reliably assess creativity. In other

words, although the possibility of a single, sufficiently rich interaction can capture the creativity is not excluded, a certain amount of “playful repetition,” and a feeling of challenge is most likely needed in order to reliably assess creativity at a single player level (El-Den, Schneider, Mirzaei, & Carter, 2020). This aligns well with how construct validity is inferred from gradually accumulated data by a diversity of approaches (Messick, 1987). Taking the recommendation for multiple tasks one step further, if properly managed, robust data sets from *suites* of games could glean vastly more information on the complex process of creativity than any one individual game or task alone. For instance, *crea.blender* belongs to a suite of games that has been designed in response to the call for portfolio-based assessment of creativity (Rafner, 2021; Rafner et al., 2020). A major challenge for such a suite would be ensuring sufficient engagement and motivation so that it does not induce testing fatigue and player disengagement (Lieberoth, Pedersen, & Sherson, 2015).

Out of the reviewed papers, only four games implemented progression (modal-, level-, or round- based) explicitly in their design. This can be taken as an indication of the “prototype” level of many of the examined games. The games without progression of any kind simply presented the player with a task and ended either when the task was completed or the time ran out (Hart et al., 2017; Krebs et al., 2020; Rozenkrantz et al., 2017). In contrast, *Physics Playground* (Shute & Rahimi, 2021) has an extensive standard playmode with level-based progression, which scores creativity of the players using stealth assessment. Additionally, it has a level editor mode, where participants can build their own levels, which are scored by expert raters. *Crea.blender* (Rafner et al., 2020) includes three modes, all with the same core game mechanics (AI-enhanced image blending), but different goals. Each of these modes are designed to assess either DT, CT, or creative products. Similar to *Physics Playground*, the ‘creative products’ mode is envisioned to be assessed by human raters. Within the CT mode, there are three levels, each increasing in difficulty.

How about convergent validity?

Table 6 highlights games which build credibility as creativity assessment instruments through correlations with standardized tests (i.e., convergent validity). Hart et al. (2017) and Rozenkrantz et al. (2017) report statistically insignificant and very weak correlations between the *Creative Foraging Game*, $r = 0.18$ ($N = 57$), and the *Alternative Uses Test*, and conclude that their task ‘does not correlate strongly with the standard task’

(i.e., low convergent validity). Taken at face value, this could imply a failure to capture the essence of creativity. However, the authors report similarly low correlations among the variety of standard tests they employed in the study. Therefore, it can be seen that the low correlations of the game metrics with standard values could be taken as an indication of limited construct validity of the standard tasks, or, as Runco et al. (2016) suggest, as an indication that different standard tests capture different aspects of creativity. The largest validation study, Physics Playground (N = 167; Shute & Rahimi, 2021) reports similarly low convergent validity with standard tests ($r = 0.18$). BuzzWords (Howell & Veale, 2009), Arctic Quest (Chuang et al., 2015), and Immune Defense (Krebs et al., 2020) stand out with reports of higher correlation values ($r > 0.5$). This could indicate that these approaches should be particularly promising for creativity assessment. However, such an interpretation should be moderated by both the lack of correlation among standardized measures, combined with the low participant numbers observed (N < 30) in games reporting high correlation. Given all the above, these values may simply indicate that the game-based interaction has been in fact, successfully implemented to match the particular processes of the selected reference task rather than capturing the general underlying phenomenon of creativity. We therefore recommend that convergent validity be used in combination with other complementary validity measures (e.g., ecological and construct) in order to more deeply understand creativity.

Mixed methods to assess and understand creativity

All of the reviewed games in this study utilize stealth assessment approaches (either implicitly or explicitly). However, stealth assessment hinges on the ability to “(a) define the claims concerning learners’ competencies, (b) establish what represents valid evidence of a claim, and (c) determine the kind of tasks or situations that will elicit that evidence” (Shute et al., 2016, p. 2). That is, the methodology involves a thorough literature review of the competency that one wants to assess (e.g., creativity). The deep review of the literature serves as the basis for crafting a multifaceted, interconnected graphical representation of creativity. As we have discussed above, realizing such established construct validity is by no means a solved problem. This underscores the general call in this paper to not only use games to assess creativity but also to deeper understand creativity. This could potentially be achieved by combining well-structured tasks tapping into CT and DT behavior and more free open-ended exploration such as in the final modes of *crea- blender* and *Physics Playground*. The open-ended

responses or products are typically not amenable to direct algorithmic scoring and have therefore traditionally been subject to expert consensual assessment. This methodology represents a major roadblock for the scaling up of these kinds of initiatives.

Building off of the literature around consensual assessment, crowdsourced peer rating of in-game produced products may present a solution to this problem. When gameplay is simple enough that anyone can learn within a relatively short period of time to distinguish various characteristics of a set of solutions/products, crowdsourced peer assessment could be a promising method for scalable creative product assessment. An example of implementation of consensual assessment based on peer feedback could be asking a player to rank on a Likert scale how creative is a product (or solution) produced by another player. It is recommended that such assessment comes after a player has completed a tutorial/introduction and has tried the game themselves, so that they are aware of how difficult/rare it is to make certain products. If crowdsourced, these products could have many peer raters, with a subset of products cross-validated by expert raters. Crowdsourced assessment could be employed as long as wording at the point of creation is identical to wording at the point of rating (e.g., “Make a creative shape,” and “Which shape is more creative?,” respectively). Notably, peer assessment was not addressed in any of the reviewed papers. Combining creativity information from products via crowdsourcing and processes via stealth assessment could possibly result in stronger construct validity of an instrument.

Game-based Convergent Thinking assessment warrants more attention

As DT and CT are interwoven during creative cognition, in many creativity assessments it may be difficult to distinguish between the two constructs based on task performance (Cortes et al., 2019). In other words some game tasks may tap into both convergent and divergent processes. Generally speaking, games with an emphasis on solving concrete challenges with a well-defined correct solution seem to lend themselves well to capturing the CT process (selection of and iteration on correct solutions). This is particularly relevant as developing a robust, and non-verbal assessment for CT is one of the bigger challenges in the field. Notably, CT was under-explored compared to DT in the reviewed games: two games assessing CT (Howell & Veale, 2009; Rafner et al., 2020) versus nine explicitly assessing DT (Ali et al., 2019; Chuang et al., 2015; Gal et al., 2017; Hart et al., 2017; Krebs et al., 2020; Rafner et al., 2020; Rozenkrantz et al., 2017;

Shillo et al., 2019; Shute & Rahimi, 2021; Voiskounsky et al., 2017). As discussed in the introduction, this bias toward DT assessment is common in current creativity research. Crea.blender (Rafner et al., 2020) describes a visual CT mode, in which players are presented with a target image and three sets of starting images. The players must choose which set produced the target image, and then recreate it. In contrast to BuzzWords (Howell & Veale, 2009; 2010; Howell & Veale, 2009), this task does not rely on the player's verbal skills for finding the right solution. However, it relies heavily on the player's ability to understand and work with the mechanisms of the image-generating algorithm, otherwise the player could not make a correct prediction. The authors addressed this exact issue in a small pilot study, and the results suggest that participants are able to understand and control the mechanics successfully (Rafner et al., 2020), based on which, this could be a promising approach for non-verbal CT assessment. The Creativity Game Based Evaluation System (Yeh et al., 2016) employs insight problems which can measure elements of both CT and DT; however, the particular study only collects data on the correctness of the responses, not the process itself. With appropriate adaptations, insight problems could potentially be a promising direction for achieving non-verbal CT assessment. However, the importance of capturing the process in addition to the correctness of the response should be underscored in future studies.

A call for interdisciplinary projects and dialogs

In order to shed more light on the scientific disciplines from which the selected papers originated, a count was made of how many of papers had at least one coauthor from the following disciplines: computer science /engineering (8), other natural science (4), education (6), psychology (8), and other (4). Of course, disciplinary boundaries can vary and not all affiliations were specified fully in the papers, meaning that the reported numbers were in some cases qualitative estimates. Still, accepting these classifications and performing a slight aggregation, an even split can be observed, following the traditional division: education/psychology (10), all natural science (10). Interestingly, only four papers were assessed to have representatives of both of these traditions. It can be speculated that the field of game-based creativity assessment has so far primarily been investigated in relatively narrowly defined scientific contexts. This could explain the generally low

participation numbers reported above. Any long-term, scalable creativity assessment game could greatly benefit from in-depth interdisciplinary communication and collaboration.

Additionally, none of the papers found during this review were published in the Creativity Research Journal, or Journal of Creative Behavior which “publish the most studies on creativity and were cited the most in the field” (Long, Plucker, Yu, Ding, & Kaufman, 2014 as seen in Long, 2014) or Philosophy, Esthetics, Psychology and the Arts which is a relatively new but prominent journal in the field (Long, 2014). This highlights a need for facilitating interdisciplinary dialogs between fundamental research on the constructs of creativity, assessment considerations, and those designing new innovative assessment instruments, which is by no means an easy task. However, it is a goal of this paper to encourage greater cooperation between those conducting fundamental research on creativity, its assessment, and those designing new innovative assessment instruments. These collaborations are necessary to address the grand challenge of achieving large-scale, psychometrically stringent creativity assessment.

Conclusions and future directions

This article set out to review how digital games have been used to assess creativity and in particular investigate if they are already or could potentially be used to combine scalability of process-oriented assessments with the more ecologically valid but much less scalable product-oriented assessments of creativity. The games presented in this paper not only span a range of disciplines and experimental settings, but also a range of game design decisions, research approach and data analysis. The review portrays a clear interest in and potential for digital games to assess creativity. However, the small number of studies indicates the infancy of the field. Many of the studies in this review have been conducted by natural *or* social science teams, which allows for speculations that at this nascent stage, the field is dominated by explorative and domain specific endeavors rather than displaying the strong interdisciplinary dialogue between computer and social sciences, which could be an important step to reach scalability. Scalability is one element that is crucial both to *understand* creativity and to subsequently *assess* it. As projects grow in depth and scale, another challenge emerges: it will be unfeasible to perform expert consensual assessment of in-game produced products. Further efforts to algorithmic assessment of such products as well as explorations

into crowdsourced peer consensual assessment should thus be investigated. Promising trends have been identified such as automated assessment using e.g., Bayesian Networks (Shute et al., 2021) and semantic-based natural language processing algorithms (Beatty & Johnson, 2021) as well as in-task AI-assistance to ideally facilitate the creation of unique products without relying on prior domain-specific skills or excessive time commitment (Rafner et al., 2020).

Although the games as a whole cover both procedural creativity elements like CT and DT as well as modes that allow for open ended solution search and/or product generation, most studies have prioritized systematic comparisons with traditional (DT) tasks. As established construct validity is by no means a solved problem, convergent validity efforts cannot stand alone. This review therefore highlights the potential of digital games to help build a sufficiently solid methodological foundation of the increasingly complex interactions to foster novel insights into the nature of creativity. To obtain this, various game modes should ideally be aggregated to a comprehensive suite of games exhibiting a diversity of interaction forms and contexts.

In conclusion, many important steppingstones toward scalable game-based assessment have been identified in the reviewed papers. At the same time, this review also highlights many of the unsolved challenges that lie ahead.

Notes

1. Ecological validity refers to the extent to which data collected from an experiment reflects real-life behavior, and the extent to which findings are generalizable to real-life settings. This is dependent on the testing environment (Andrade, 2018).
2. Construct validity refers to the extent to which an instrument is able to capture the phenomenon/ability it is trying to measure (Cronbach & Meehl, 1955).
3. Convergent validity is a type of construct validity, which refers to the extent to which two tests that theoretically should be related, are in fact correlated (Cronbach & Meehl, 1955).
4. Internal validity refers to the extent to which the study is able to answer the research question, without any bias, given the study design, the conduct, and analysis (Andrade, 2018).
5. Game *variants* have specialized prompts for the academic study but the game itself has not been altered.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Carlsbergfondet; Synakos Foundation; Novo Nordisk Fonden.

ORCID

Janet Rafner  <http://orcid.org/0000-0001-9264-3334>

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

References

- Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, 26(2), 229–238. doi:10.1080/10400419.2014.901095.
- Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 153. doi:10.1037/aca0000231.
- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, 15(2), 163.
- Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, 47(4), 823–837. doi:10.1177/0149206320969787.
- Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018, July). *Improving stealth assessment in game-based learning with LSTM-based analytics*. International Conference on Educational Data Mining.
- Ali, S., Moroso, T., & Breazeal, C. (2019). Can children learn creativity from a social robot? *Proceedings of the 2019 on Creativity and Cognition*. doi:10.1145/3325480.3325499.
- Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives*, 12(1–2), 1–33. doi:10.1080/15366367.2014.910060.
- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013. doi:10.1037/0022-3514.43.5.997.
- Andrade, C. (2018). Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine*, 40(5), 498–499. doi:10.4103/IJPSYM.IJPSYM_334_18.
- Bachelor, P. A. (1989). Maximum likelihood Confirmatory Factor-analytic investigation of factors Within Guilford's structure of INTELLECT MODEL. *Journal of Applied Psychology*, 74(5), 797–804. doi:10.1037/0021-9010.74.5.797.
- Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, 16(1), 113–117. doi:10.1207/s15326934crj1601_11.
- Barbot, B., & Reiter-Palmon, R. (2019). Creativity assessment: Pitfalls, solutions, and standards. *Psychology Of Aesthetics, Creativity, And The Arts*, 13(2), 131–132. doi:10.1037/aca0000251.

- Beatty, 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.
- Beck, S., Bergenholtz, C., Bogers, M., Brasseur, T.-M., Conradsen, M. L., Di Marco, D., ... Xu, S. M. (2020). The Open Innovation in Science research field: A collaborative conceptualization approach. *Industry and Innovation*, 1–50. doi:10.1080/13662716.2020.1792274.
- Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology*, 12(2), 210. doi:10.5964/ejop.v12i2.1127.
- Belsky, S. (2020). *Creativity will be key to competing against AI in the future workforce - here's how*. World Economic Forum. Retrieved April 14, 2021, from <https://www.weforum.org/agenda/2020/11/ai-automation-creativity-workforce-skill-fute-of-work/>
- Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, 35(4), 634–639. doi:10.3758/bf03195543.
- Carson, S. (2006, April 19). *Creativity and mental illness*. Invitational Panel Discussion Hosted by Yale's Mind Matters Consortium, New Haven, CT.
- Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological assessment: Past and future. *Journal of the International Neuropsychological Society: JINS*, 23(9–10), 778. doi:10.1017/S1355617717001060.
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in human behavior*, 29(6), 2156–2160. Chicago
- Chuang, T. Y., Liu, E. Z. F., & Shiu, W. Y. (2015). Game-based creativity assessment system: The application of fuzzy theory. *Multimedia Tools and Applications*, 74(21), 9141–9155. doi:10.1007/s11042-014-2070-7.
- Clapham, M. (1996). The construct validity of divergent scores in the structure-of-intellect learning abilities test. *Educational And Psychological Measurement*, 56(2), 287–292. doi:10.1177/0013164496056002009.
- Cooper, E. (1991). A critique of six measures for assessing creativity. *The Journal of Creative Behavior*, 25(3), 194–204. doi:10.1002/j.2162-6057.1991.tb01370.x.
- 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.
- Cotter, K. N., & Silvia, P. J. (2019). Ecological assessment in research on aesthetics, creativity, and the arts: Basic concepts, common questions, and gentle warnings. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 211–217. doi:10.1037/aca0000218.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281. doi:10.1037/h0040957.
- Cropley, A. (2006). In Praise of Convergent Thinking. *Creativity Research Journal*, 18(3), 391–404. doi:10.1207/s15326934crj1803_13.
- Cseh, G., & Jeffries, K. (2019). A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. *Psychology Of Aesthetics, Creativity, And The Arts*, 13(2), 159–166. doi:10.1037/aca0000220.
- Csikszentmihalyi, M. (1997). Flow and the psychology of discovery and invention. *HarperPerennial*, New York, 39.
- DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction*, 2012, 1–7. doi:10.1155/2012/654791.
- DeYoung, C. G., Flanders, J. L., & Peterson, J. B. (2008). Cognitive abilities involved in insight problem solving: An individual differences model. *Creativity Research Journal*, 20(3), 278–290. doi:10.1080/10400410802278719.
- DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, 17(1), 17–28.
- Dontcheva, M., Gerber, E., & Lewis, S. (2011, May). *Crowdsourcing and creativity*. CHI 2011: crowdsourcing workshop (Vol. 10, p. 4).
- Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem–solution. *Design Studies*, 22(5), 425–437. doi:10.1016/S0142-694X(01)00009-6.
- Dow, G. T., & Mayer, R. E. (2004). Teaching students to solve insight problems: Evidence for domain specificity in creativity training. *Creativity Research Journal*, 16(4), 389–398. doi:10.1080/10400410409534550.
- El-Den, S., Schneider, C., Mirzaei, A., & Carter, S. (2020). How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, 28(4), 326–336.
- Evans, N. J., Holmes, W. R., & Trueblood, J. S. (2019). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychonomic Bulletin & Review*, 26(3), 901–933. doi:10.3758/s13423-018-1557-z.
- Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: Computer applications* (pp. 231–243). In Roberto PoliMichael Healy Achilles Kameas (Eds.), Dordrecht, Netherlands: Springer.
- Figment. Lumos Labs, I. (2019, July 16). *Figment by LUMOSITY*. App Store. <https://apps.apple.com/ca/app/figment-by-lumosity/id1142719997>
- Forthmann, B., Oyeade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, 53(4), 559–575. doi:10.1002/jocb.240.
- Gal, L., Hershkovitz, A., Morán, A. E., Guenaga, M., & Garaizar, P. (2017). Suggesting a log-based creativity measurement for online programming learning environment. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. doi:10.1145/3051457.3054003.
- Gardner, H. (1988). Creativity: An interdisciplinary perspective. *Creativity Research Journal*, 1(1), 8–26. doi:10.1080/10400418809534284.
- Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-learning and Digital Media*, 2(1), 5–16. doi:10.2304/elea.2005.2.1.5.
- Ghuman, D., & Griffiths, M. (2012). A cross-genre study of online gaming: Player demographics, motivation for play, and social interactions among players. *International Journal of Cyber Behavior, Psychology and Learning (IJCBL)*, 2(1), 13–29. doi:10.4018/ijcbl.2012010102.
- Green, G., & Kaufman, J. C. (2015). *Video games and creativity*. Academic Press. <https://www.elsevier.com/books/video-games-and-creativity/green/978-0-12-801462-2>

- Grey, A. (2016). *The 10 skills you need to thrive in the Fourth Industrial Revolution*. World Economic Forum. Retrieved April 14, 2021, from <https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/>
- Guilford, J. (1950). Creativity. *American Psychologist*, 5(9), 444–454. doi:10.1037/h0063487.
- Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, 53(4), 267. doi:10.1037/h0040755.
- Guilford, J. P. (1968). *Intelligence, creativity, and their educational implications*. Edits Pub, University of Michigan.
- Guilford, J. P. (1973). Characteristics of Creativity.
- Guo, J. (2019). Web-based creativity assessment system that collects both verbal and figural responses: Its problems and potentials. *International Journal of Information and Education Technology*, 9(1), 27–34. doi:10.18178/ijiet.2019.9.1.1168.
- 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.
- Howell, S., & Veale, T. (2009). *Designing serious games with linguistic resources*. Proceedings of the 4th International Conference on Foundations of Digital Games - FDG '09. doi:10.1145/1536513.1536564.
- Howell, S., & Veale, T. (2010). Assessing Verbal Creativity with Digital Educational Games. In proceedings if of the 2010 International Conference on Engaging Pedagogy
- Huber, K. J., & Oded, N., Sharpe Wessling. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, 44(1), 211–230.
- Hyde, M., Scott-Slade, M., Scott-Slade, H., Hornberger, M., Spiers, H., Dalton, R., ... Bohbot, V. (2016). Sea Hero Quest: The World's first mobile game where anyone can help scientists fight dementia.
- Indeed Editorial Team. (2021). *Skill assessment tests: Definition and examples*. Indeed.com. Retrieved April 14, 2021, from <https://www.indeed.com/career-advice/career-development/skills-assessment-test>
- Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22(3), 374–419. doi:10.1016/0010-0285(90)90008-R.
- Katz-Buonincontro, J., & Anderson, R. C. (2018). A review of articles using observation methods to study creativity in education (1980–2018). *The Journal of Creative Behavior*, 54(3), 508–524. doi:10.1002/jocb.385.
- Kaufman, J. (2019). Self-assessments of creativity: Not ideal, but better than you think. *Psychology Of Aesthetics, Creativity, And The Arts*, 13(2), 187–192. doi:10.1037/aca0000217.
- Kaufman, J. C., & Sternberg, R. J. (Eds.). (2010). *The Cambridge handbook of creativity*. Cambridge University Press
- Kaufman, J. C., & Sternberg, R. J. (Eds.). (2019). *The Cambridge handbook of creativity* (2nd ed.). Cambridge University Press.
- Keefe, J. A., & Magaro, P. A. (1980). Creativity and schizophrenia: An equivalence of cognitive processing. *Journal of Abnormal Psychology*, 89(3), 390. doi:10.1037/0021-843X.89.3.390.
- Kettlemind. (n.d.) Kettlemind. <https://www.kettlemind.com/>
- Kittur, A. (2010). Crowdsourcing, collaboration and creativity. *XRDS: Crossroads, the ACM Magazine for Students*, 17(2), 22–26. doi:10.1145/1869086.1869096.
- KnackApp. (n.d.) *Join KnackApp*. KnackApp. <https://knackapp.com/>
- Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D. A., Bick, M. J., ... Baker, D. (2019). De novo protein design by citizen scientists. *Nature*, 570(7761), 390–394. doi:10.1038/s41586-019-1274-4.
- Krebs, E., Jaschek, C., von Thienen, J., Borchart, K.-P., Meinel, C., & Kolodny, O. (2020). *Designing a Video Game to Measure Creativity*. 2020 IEEE Conference on Games (CoG). doi:10.1109/cog47356.2020.9231672.
- Kühn, S., Ritter, S. M., Müller, B. C., Van Baaren, R. B., Brass, M., & Dijksterhuis, A. (2014). The importance of the default mode network in creativity—A structural MRI study. *The Journal of Creative Behavior*, 48(2), 152–163. doi:10.1002/jocb.45.
- Kwon, M., Goetz, E. T., & Zellner, R. D. (1998). Developing a computer-based TTCT: Promises and problems. *The Journal of Creative Behavior*, 32(2), 96–106. doi:10.1002/j.2162-6057.1998.tb00809.x.
- Lau, S., & Cheung, P. C. (2010). Creativity assessment: Comparability of the electronic and paper-and-pencil versions of the Wallach–Kogan Creativity Tests. *Thinking Skills and Creativity*, 5(3), 101–107. doi:10.1016/j.tsc.2010.09.004.
- Long, H. (2014). An empirical review of research methodologies and methods in creativity studies (2003–2012). *Creativity Research Journal*, 26(4), 427–438. doi:10.1080/10400419.2014.961781.
- Long, H., Plucker, J. A., Yu, Q., Ding, Y., & Kaufman, J. C. (2014). Research productivity and performance of journals in the creativity sciences: A bibliometric analysis. *Creativity Research Journal*, 26(3), 353–360. doi:10.1080/10400419.2014.929425.
- Lumosity. (n.d.) *Lumosity brain training: Challenge & improve your mind*. Lumosity. <https://www.lumosity.com/en/>
- Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games*, 4(2), e11. doi:10.2196/games.5888
- Maheshwari, A. K., & Werd, M. R. P. (2019). Architecture and creativity: Examining the impact of Maharishi Vastu on workplace creativity. *Creativity Research Journal*, 31(4), 371–376. doi:10.1080/10400419.2019.1667943.
- Matud, M. P., Rodríguez, C., & Grande, J. (2007). Gender differences in creative thinking. *Personality and Individual Differences*, 43(5), 1137–1147. doi:10.1016/j.paid.2007.03.006.
- 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.
- McKenna, H. P., Arnone, M. P., Kaarst-Brown, M. L., McKnight, L. W., & Chauncey, S. A. (2013). *Application of the consensual assessment technique in 21st century technology-pervasive learning environments*. In Proceedings of the 6th International Conference of Education, Research and Innovation (iCERi2013) (pp. 6410–6419).

- Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, 69(3), 220–232. doi:10.1037/h0048850.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115. doi:10.1086/288135.
- Messick, S. (1987). Validity. *ETS Research Report Series*, 1987(2), i–208. doi:10.1002/j.2330-8516.1987.tb00255.x.
- Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., et al. (2015, June). DeepStealth: Leveraging deep learning models for stealth assessment in game-based learning environments. In International conference on artificial intelligence in education (pp. 277–286). Cham, Switzerland: Springer.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1), i–29. doi:10.1002/j.2333-8504.2003.tb01908.x.
- Noy, L., Hart, Y., Andrew, N., Ramote, O., Mayo, A. E., & Alon, U. (2012, May). A quantitative study of creative leaps. In ICCO (pp. 72–76). International Conference on Computational Creativity Dublin, Ireland.
- Oltețeanu, A.-M., & Zunjani, F. H. (2020). Frontiers in psychology. *A Visual Remote Associates Test and Its Validation*, 11. doi:10.3389/fpsyg.2020.00026.
- Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., & Hosio, S. (2020, April). *Creativity on paid crowdsourcing platforms*. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14).
- Palaniappan, A. K. (2012). Web-based creativity assessment system. *International Journal of Information and Education Technology*, 255–258. doi:10.7763/ijiet.2012.v2.123.
- Pásztor, A., Molnár, G., & Csapó, B. (2015). Technology-based assessment of creativity in educational context: The case of divergent thinking and its relation to mathematical achievement. *Thinking Skills and Creativity*, 18, 32–42. doi:10.1016/j.tsc.2015.05.004.
- 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. *arXiv Preprint arXiv:2009.05274*.
- Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, 39(2), 83–96. doi:10.1207/s15326985ep3902_1.
- Rafner, J. (2021). Creativity assessment games and crowdsourcing. In *Creativity and Cognition* (pp. 1–5).
- Rafner, J., Hjorth, A., Risi, S., Philipsen, L., Dumas, C., Biskjaer, M. M., ... Sherson, J. (2020). *crea.blender*. Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play. doi:10.1145/3383668.3419907.
- Rahimi, S., Shute, V., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, 5(2), 141–165. doi:10.46328/ijtes.118.
- Rahimi, S., & Shute, V. J. (2021). The effects of video games on creativity: A systematic review. In *Handbook of lifespan development of creativity* (pp. 37). Cambridge: Cambridge University Press.
- 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.
- Remy, C., MacDonald Vermeulen, L., Frich, J., Biskjaer, M. M., & Dalsgaard, P. (2020, July). *Evaluating creativity support tools in HCI research*. Proceedings of the 2020 ACM Designing Interactive Systems Conference (pp. 457–476).
- Rozenkrantz, L., Mayo, A. E., Ilan, T., Hart, Y., Noy, L., & Alon, U. (2017). Placebo can enhance creativity. *PLOS ONE*, 12(9), e0182466. doi:10.1371/journal.pone.0182466.
- 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., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96. doi:10.1080/10400419.2012.650092.
- Sawyer, R. K. (2011). *Explaining creativity: The science of human innovation*. Oxford university press.
- Schleicher, A. (n.d.). *The case for 21st-century learning - OECD*. Oecd.org. Retrieved April 14, 2021, from <https://www.oecd.org/general/thecasefor21stcenturylearning.htm>
- Schram, A., & Ule, A. (Eds.). (2019). *Handbook of research methods and applications in experimental economics*.
- Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (Eds.). (2019). *A handbook of process tracing methods*. Routledge.
- Shillo, R., Hoernle, N., & Gal, K. (2019). Detecting creativity in an open ended geometry environment. *International Educational Data Mining Society*.
- Shute, V., & 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.
- Shute, V., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. *The Handbook of Cognition and Assessment*, 535–562. doi:10.1002/9781118956588.ch22.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2), 503–524.
- Shute, V. J., Lu, X., & Rahimi, S. (2021). Stealth assessment. In J. M. Spector (Ed.), *The Routledge Encyclopedia of Education* (pp. 1–9). London, UK: Taylor & Francis group.
- Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. Michael Spector, M. David Merrill, Jan Elen, M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 311–321). New York, NY: Springer.
- Shute, V. J., Lu, X., & Rahimi, S. (2021). Stealth assessment. In J. M. Spector (Ed.), *The Routledge Encyclopedia of Education* (pp. 1–9). London, UK: Taylor & Francis group.
- Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 296).
- Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, Massachusetts: The MIT Press.

- Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, 80, 58–67. doi:10.1016/j.compedu.2014.08.013.
- Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63, 106–117. doi:10.1016/j.chb.2016.05.047.
- Strobach, T., & Kupferberg, A. (2020). Commercial Brain Training. *Cognitive Training*, 289–305.
- Stuart, D. (2009). Assessment instruments for the global workforce. In *Contemporary leadership and intercultural competence: Exploring the cross-cultural dynamics within organizations* (pp. 175–190).
- Thompson, B., & Andersson, B. V. (1983). Construct validity of the divergent production subtests from the structure-of-intellect learning abilities test. *Educational and Psychological Measurement*, 43(2), 651–655. doi:10.1177/001316448304300240.
- Torrance, E. P. (1966). The torrance tests of creative thinking-norms-technical manual research edition-verbal tests. *Forms A and B-Figural Tests, Forms A and B*.
- Torrance, E. P. (1976). Creativity testing in education. *Creative Child & Adult Quarterly*, 1(3), 136–148.
- Torrance, E. P. (1981). *Thinking creatively in action and movement*. Bensenville, IL: Scholastic Testing Service.
- Torrance, E. P. (1988). The nature of creativity as manifest in its testing. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 43–75). New York: Cambridge University Press.
- Valladares-Rodríguez, S., Pérez-Rodríguez, R., Anido-Rifón, L., & Fernández-Iglesias, M. (2016). Trends on the application of serious games to neuropsychological evaluation: A scoping review. *Journal of Biomedical Informatics*, 64, 296–319. doi:10.1016/j.jbi.2016.10.019.
- Voiskounsky, A. E., Yermolova, T. D., Yagolkovskiy, S. R., & Khromova, V. M. (2017). Creativity in online gaming: Individual and dyadic performance in Minecraft. *Psychology in Russia: State of the Art*, 10(4), 144–161. doi:10.11621/pir.2017.0413.
- Wadum, A. L., Pedersen, M. K., & Sherson, J. (2015). Play or science?: a study of learning and framing in crowdsience games. Well Played
- Wallach, M. A., & Kogan, N. (1965). Modes of thinking in young children.
- Williams, F. E. (1980). *Creativity assessment packet*. Buffalo, NY: DOK Publishers.
- World Economic Forum. (2020). *Future of Jobs Report*. <https://www.weforum.org/reports/the-future-of-jobs-report-2020>
- Yeh, Y.-C., Lai, S. C., & Lin, C.-W. (2016). The dynamic influence of emotions on game-based creativity: An integrated analysis of emotional valence, activation strength, and regulation focus. *Computers in Human Behavior*, 55, 817–825. doi:10.1016/j.chb.2015.10.037
- Zojceska, A. (2019). *20 Reasons to Use Skill Assessment Tests*. TalentLyft. Retrieved April 14, 2021, from <https://www.talentlyft.com/en/blog/article/277/20-reasons-to-use-skill-assessment-tests>

Appendix A

Screenshots from the 11 games, with original figure titles.



Figure 2 Creatures and Open Play mode interface. Top row: six source images that players can blend together. Middle row: last generated image. Bottom row: images that the player has saved.

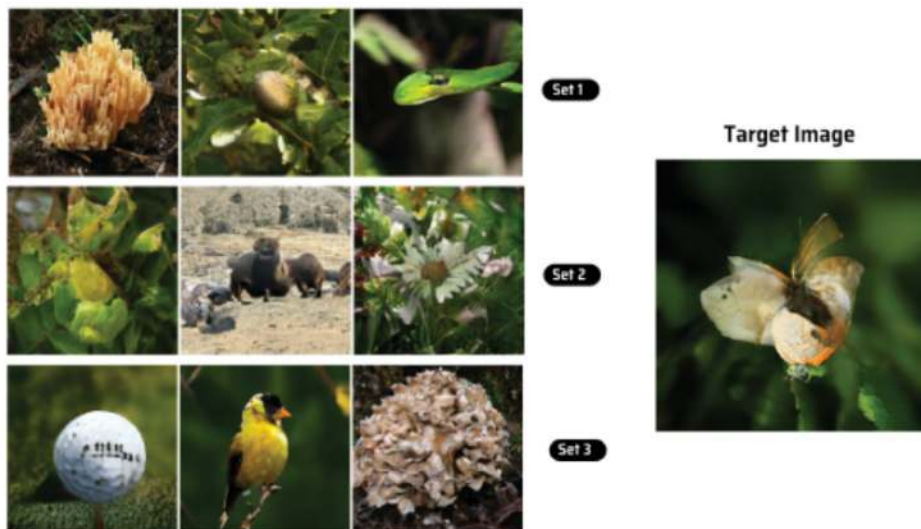


Figure 3 Challenge mode interface. The source image sets are on the left and the target image is on the right. When a set is chosen, the player blends them similar to the other modes (Figure 2)

Figure A1. Crea.blender (Rafner et al., 2020).

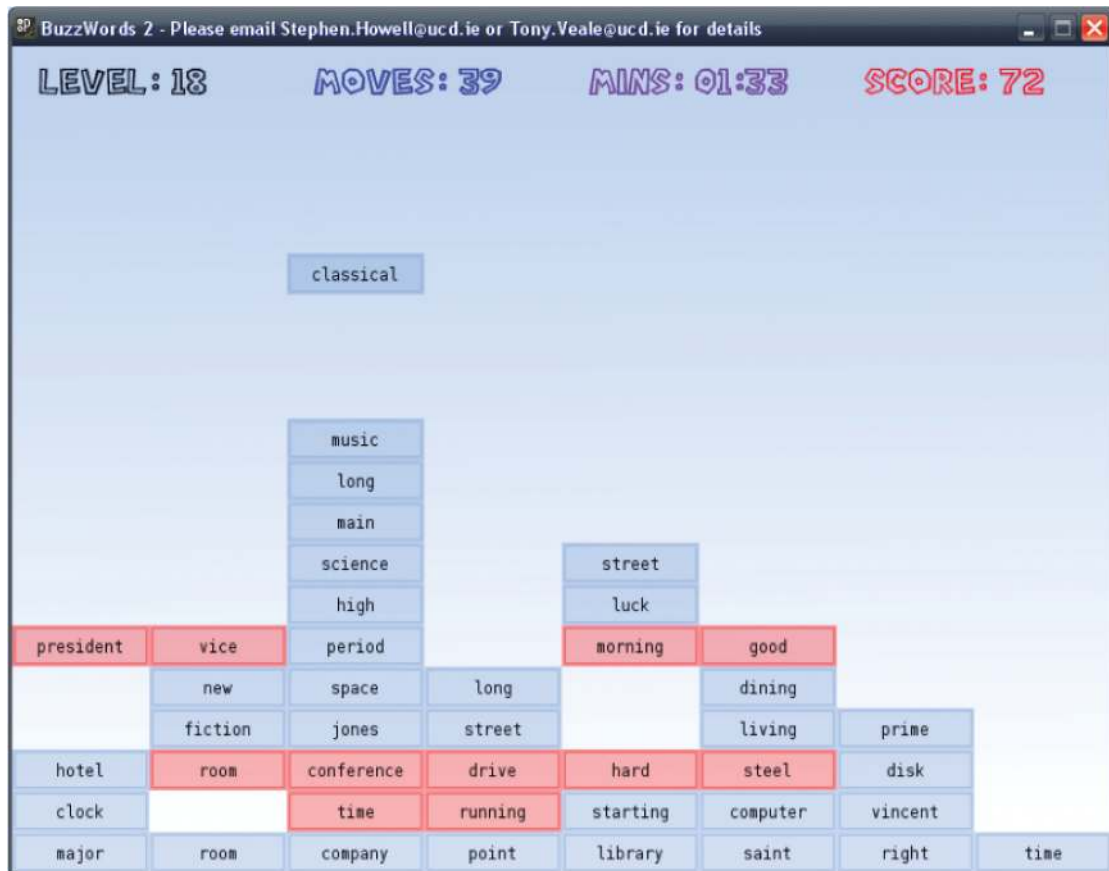


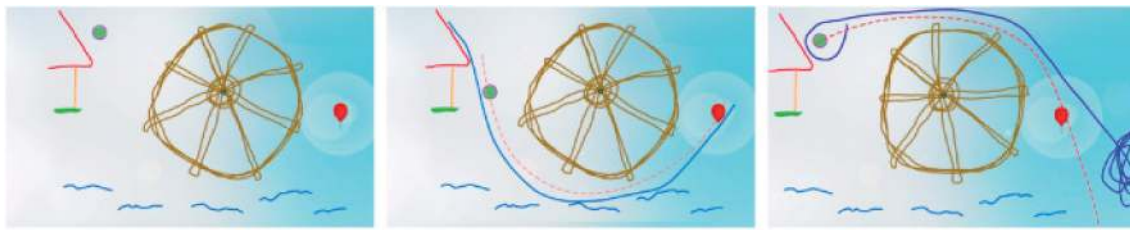
Figure 1: BuzzWords in action, note locked words are in red, *classical* is in play and will connect with *music* to score

Figure A2. BuzzWords (Howell et al., 2009).



Figure 2. (a) Interaction Scene. A child is playing the Doodle Creativity Game on an Android Tablet with the social robot Jibo. (b) Example of a Doodle Image. 10 Doodles were used in the Doodle Creativity Game (5 per player).

Figure A3. Doodle (Ali et al., 2019).



2a 2b 2c
Fig. 2. The *Big Watermill* level (2a) and two possible solutions (2b and 2c).

Figure A4. Physics Playground (Shute et al., 2021).



Figure 1. Level 9 of Kodetu. The user should edit a code, using the available blocks, to get the astronaut to the red mark (the coding-area is not shown here)

Figure A5. Kodetu (Gal et al., 2017).



Arctic Quest 2: Trophy Room

Figure A6. Arctic Quest 2 (Chuang et al., 2015).

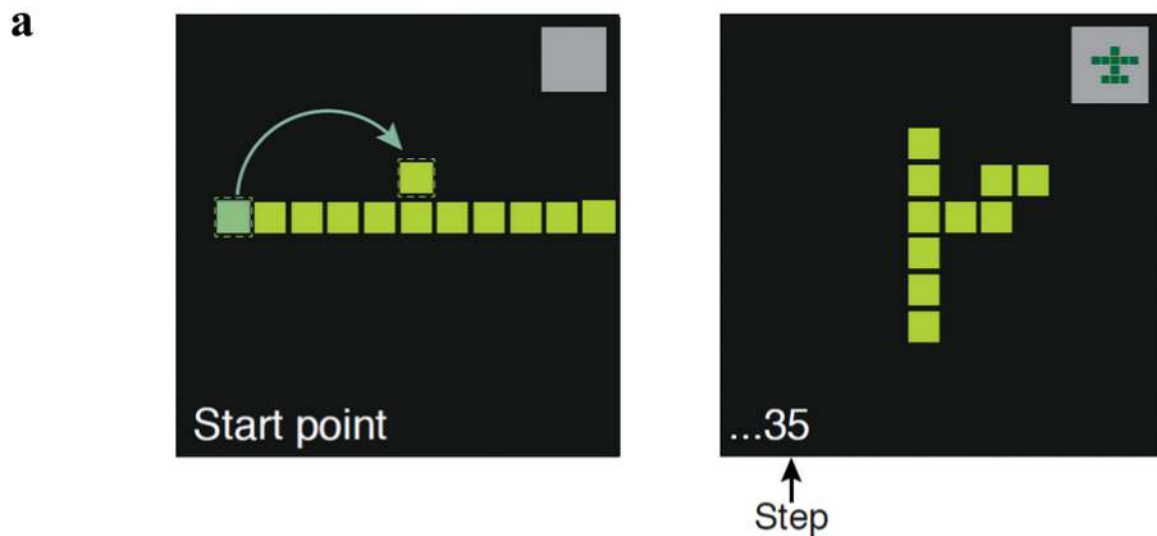


Fig 1. Creative foraging game (CFG). a: CFG interface. Left: starting point; right: example of a shape.

Figure A7. Creative Foraging Game (Rozenkrantz et al., 2017).

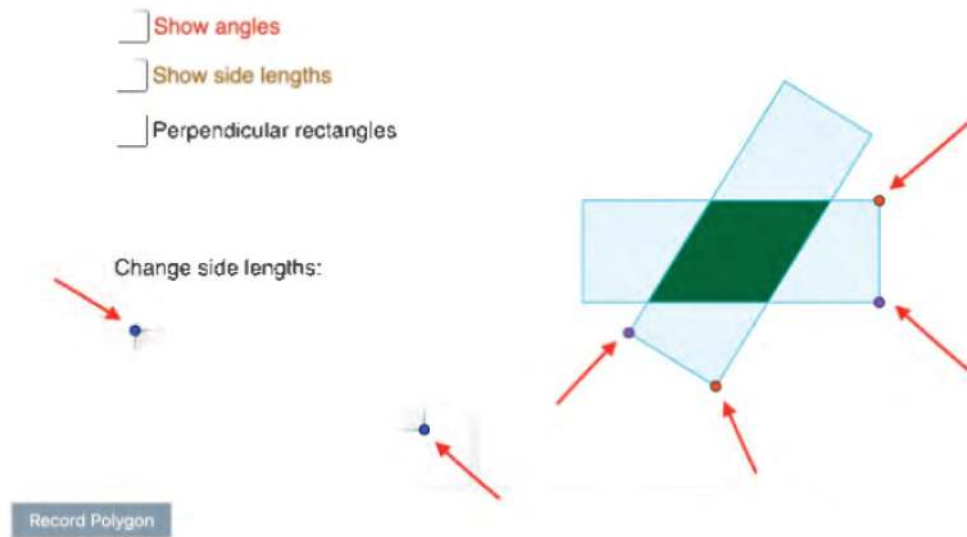


Figure 1: Open Creativity App

Figure A8. Open Creativity App (Shillo et al., 2019).

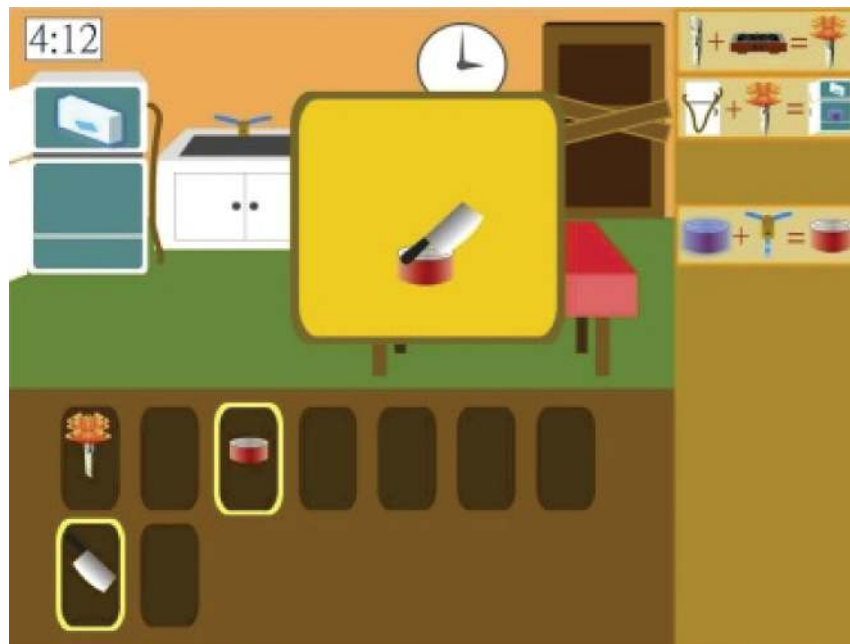


Fig. 2. An example screen in the game.

Figure A9. Creativity Game-Based Evaluation System (Yeh et al., 2016).

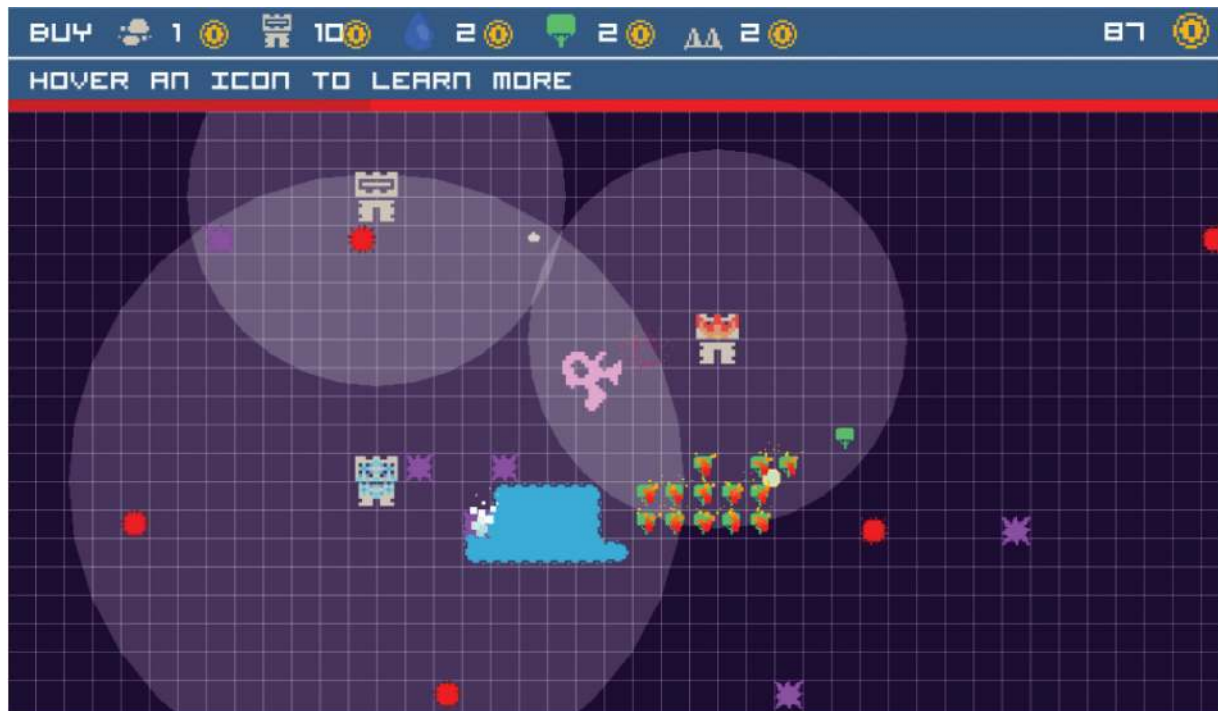


Fig. 1. A screenshot of a game in progress. In the center, the heart can be seen, which needs to be protected from enemies. The player has already sustained some damage, which caused the heart to break apart. The top menu allows the players to see how much money they currently have to buy upgrades. Two types of enemies are currently attacking. The player has placed three towers, the right-most tower being upgraded to shoot fire antibodies, which have set fire to the forest. Enemies passing through the forest will also catch fire. The bottom-most tower has been upgraded to shoot ice projectiles that froze a wet enemy, hindering that enemy from advancing.

Figure A10. Immune Defense (Krebs et al., 2020).

Appendix B Additional tables

Table B1. Game instructions.

Game	Instructions
crea.blender (Rafner et al., 2020, p. 341)	(1) Creatures: (Figure 2) Players are presented with six images and are asked to create and save as many different “animal-like” figures as possible in five minutes. (2) Challenge: (Figure 3) Players are presented with a target image and three sets of three source images. Only one set can produce the exact target image, and players’ objective is first to determine which set was used to create the target image (up to 30 seconds), and then to recreate their closest approximation of the target image (up to three min). There are three levels in the Challenge mode. (3) Open Play: (Figure 2) Players are presented with the same six source images as they used in Creatures mode. However, now they are asked to create any image (not just animal-like) they find interesting during five minutes of playtime.
Doodle (Ali et al., 2019, pg. 32)	The robot explains the gameplay to the participants: ‘I have a fun game for you today. Do you want to play with me? Do you like telling stories? We will look at some pictures, and try to think of funny and creative ideas about what they are. Let your imagination run wild. Come up with as many ideas as you can in 30 seconds. Let me show you an example.’ The tablet app displays an example doodle selected from our library of Doodles, and the robot responds with a Doodle title. After the example Doodle, the robot asks, ‘Are you ready to start the game?’ The participants can respond with a ‘Yes’ or a ‘No’ both through speech and through buttons on the robot’s screen. If the participants respond with a ‘Yes,’ then the robot begins the game. If the participants respond with a ‘No,’ then the robot proceeds to provide a second example Doodle. If the participants respond with ‘No’ more than once, experimenters intervene to help the participants understand the game until they are ready to play.
Physics Playground (Shute & Rahimi, 2021, p. 6)	Specifically, researchers instructed participants that, “Your goal is to solve as many of the problems, in as many awesome ways as you can. The tools we taught you will come in handy for many problems. Feel free, however, to solve any problem in whatever way you like.” These instructions were provided to encourage students to both do well in the game and be creative (i.e., solve levels in as many ways as they could).
CFG (Noy et al., 2012, p. 73)	In this last stage players observe the ‘shapes gallery’ and were asked to choose ‘the five most creative shapes you discovered.’
CFG (Rozenkrantz et al., 2017, p. 3)	in which participants are asked to search for interesting and beautiful shapes in a defined geometric space
CFG (Hart et al., 2017)	The players are instructed to save shapes that they find interesting and beautiful into a gallery.
Minecraft (Voiskounsky et al., 2017, p. 151)	Please build a ship/a house which you have never seen. Try to make it as unique, un-usual, and original as possible. You can build whatever you want. There are no special requirements about the structure and shape of the construction. Activate the potential of your imagination, throw away all barriers and limitations, and just do it! Upon the completion of the task, please assign an unusual name to the constructed object. Please verbalize aloud everything you are doing at the moment, articulate all your intentions concerning the task, and—this is especially important—express verbally all the ideas you produce. Try to accomplish the mission within one hour. Upon the completion of the mission, please give a guided tour of your house (or your ship) and tell a short story about it.
Open Creativity (Shillo et al., 2019, p. 410)	All participants were requested to “create as many different polygons types as possible by intersecting two rectangles.

Table B2. Information extracted from the papers based on the codes.

Game	Creativity	Component	Duration	Participants	Location	Compensation	Availability
crea.blender (Rafner et al., 2020)	Co-creativity	CT and DT	40 m	8		No	No
BuzzWords (Howell et al., 2009)	Individual	CT	-	-	-	-	No
Doodle (Ali et al., 2019)	Co-creativity	DT	App. 10 m	48	School	No	No
Physics Playground (Almond et al., 2014)	Individual	-	-	-	-	-	No
Physics Playground (Shute et al., 2016)	Individual	DT	-	167	-	No	No
BuzzWords (Howell et al., 2010)	Individual	CT	30 m	26	-	No	No
Physics Playground (Shute et al., 2021)	Individual	DT	4 h	167	School’s computer lab	25 USD gift card	No
Kodetu (Gal et al., 2017)	Individual	DT	-	-	-	-	Yes
Arctic Quest 2 (Chuang et al., 2015)	Individual	DT	2 weeks	30			Yes
CFG (Noy et al., 2012)	Individual	Creative leaps	Max 25 m	123			No
CFG (Rozenkrantz et al., 2017)	Individual	DT	10 m	90	Research lab	10 USD	No
Minecraft (Voiskounsky et al., 2017)	Individual, dyadic	DT	20 m – 2 h	49	Online		Yes
OCA (Shillo et al., 2019)	Individual	DT	Max. 10 m	183	Online	MTurkers 10 cent/shape	No
CGES (Yeh et al., 2016)	Individual	Insight	App. 40 m	266	Laboratory	App. 10 USD	No
Immune Defense (Krebs et al., 2020)	Individual	DT	5 m	17		Sweets	No
CFG (Hart et al., 2017)	Individual	DT	15 m	100		5 USD or credit	No