Designing a Video Game to Measure Creativity

Eva Krebs*
Hasso Plattner Institute
Potsdam, Germany
eva.krebs@hpi.de

Corinna Jaschek*
Hasso Plattner Institute
Potsdam, Germany
corinna.jaschek@student.hpi.de

Julia von Thienen*
Hasso Plattner Institute
Potsdam, Germany
julia.vonThienen@hpi.de

Kim-Pascal Borchart
Hasso Plattner Institute
Potsdam, Germany
kim-pascal.borchart@student.hpi.de

Christoph Meinel
Hasso Plattner Institute
Potsdam, Germany
christoph.meinel@hpi.de

Oren Kolodny
The Hebrew University of Jerusalem
Jerusalem, Israel
oren.kolodny@mail.huji.ac.il

***Abstract*—Creativity is a central phenomenon in human life. World-famous scientists and artists are praised for their creative genius. Schools and universities seek to educate creativity in students and many employers want to hire creative personnel. However, the measurement of creativity is difficult up to the present day. Standard creativity tests typically require human expertise in the evaluation of test responses. This evaluation is often more time-intensive than taking the test itself. Moreover, creativity tests are still regularly conducted in a pen-and-paper format, rendering the data analysis all the more tedious. In this article, we propose a digital game for assessing creativity. It can be hosted online. The data analysis can be automated and conducted in real-time. The test is implemented as a tower defense game ("Immune Defense"). We submit that a video game constitutes a natural setting, as opposed to a formal testing scenario, and provides an opportunity to test real-life creativity. We use the game event data gathered during each round of the game to determine a player's creativity score. A study with 17 participants was performed to compare game-based creativity scores to scores obtained with a standard creativity test, the Alternative Uses Task. Our preliminary data validate the proposed approach.**

***Index Terms*—Alternative uses task, automation, assessment, creativity, games for research, immune defense, measurement, video game**

I. Introduction

Creativity shapes our every-day experiences profoundly. It drives developments in all areas of life, including science, technology, art, and business [1], [2]. Creativity is also a key teaching aim at schools and universities; students shall become good problem-solvers and innovators in all fields of life [3]. Similarly, many employers seek to hire staff with marked creative capacities. Thus, in academia, research, and work contexts creativity measurements are in high demand.

To this day, creativity assessments are often conducted with standardized tests in a pen-and-paper format. This approach has several drawbacks. First, administering the test itself is time- and resource-intensive, which is why only a limited number of participants can be included in pen-and-paper tests. This makes scaling studies to include a larger, possibly international audience difficult. Second, in most standard creativity tests responses of participants need to be judged by experts on dimensions such as response originality and flexibility. Evaluating a participant's test responses is typically more time-intensive than taking the test itself. Third, the non-automated rating procedure by human experts is problematic with regard to measurement reliability: a test response that one expert finds original can be considered mundane by another. Fourth, some standard creativity tests are based on self-reports. Test-takers can cheat easily by simply claiming more creative achievements than they have produced in their lives. Fifth, many creativity tests are conducted in an artificial, formal setting, and participants are asked to perform behaviors that rarely ever occur in real life, such as thinking up uncommon uses for everyday objects. Here the question of ecological validity becomes pertinent and it would seem favorable if researchers could assess creativity as occurring in real life, or more natural settings (at least, in addition to more traditional creativity assessments).

In this paper, we propose a different way to assess creativity: a video game by the name of *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 the open-source game engine Godot.[1] All game actions taken by the player (such as placing objects) as well as game events (such as an enemy getting wet after coming into contact with water) are automatically tracked and stored in a digital format. The game data can be evaluated automatically in terms of creativity scores. Such creativity scores can be generated based on the game data of one person alone, or they can be calculated based on comparisons across game data from all participants.

With the game, we intend to overcome some drawbacks of pen-and-paper tests as discussed above. The video game can be administered in highly resource-effective ways. It can even be hosted on platforms like Amazon Mechanical Turk

*Corresponding authors

[1]https://godotengine.org/

978-1-7281-4533-4/20/$31.00 ©2020 IEEE

(MTurk),[2] so that it becomes feasible to test large numbers of participants and international samples of test-takers. Game data is evaluated automatically. No human experts are needed for judgments. Thus, the data analysis is fast, it can even occur in real-time, and measurement reliability is no issue any longer. Moreover, test-takers cannot cheat in the sense that they claim greater creative capacities than they possess, because it takes real capacities to come up with creative actions in the game. Finally, using tests that approximate a natural setting has been a goal of creativity tests for many years [4]. We submit that playing a video game and avoiding a formal, test-like setting is an effective means to achieve this goal.

To ensure that the game can serve as a creativity test, we set several design goals. The game should be deterministic with no randomness introduced by the game itself, to allow reliable comparisons between different playthroughs. The gameplay itself should be systematic enough to be recorded and analyzed automatically, yet it should allow the player to experiment by providing a "combinatorial explosion" of gameplay elements. In addition to more technical goals, the game should fulfill certain usability requirements in order to be easily included in studies. Most importantly, the game needs to be easy to understand, which is why we chose an established game genre (tower defense) and added an in-game tutorial. The game should not include sensitive topics that may distract players. At the same time, the game goal or other game features should induce a sense of urgency in the player, to ensure player engagement.

We hypothesize that our game creativity scores will map to creativity scores from established creativity tests, such as the Alternative Uses Task.

In this paper, we first provide some background information on creativity models and creativity in video games in section II. Section III describes our design goals in detail, as well as the gameplay, and measurements taken. Section IV provides an overview of the study that was conducted, followed by results in section V and a discussion of findings as well as an outlook towards future work in section VI.

II. Background

In this section, we provide a brief overview of how creativity is commonly defined and measured. We also review games as a medium for creativity.

A. Creativity Theory

According to the standard definition, a solution is creative when it is novel and effective [1], [5].

Novelty means that a solution is uncommon, original, unique. Novelty can be assessed on individual and aggregate levels [6]–[8]. When a person comes up with a so-called "P-creative" solution, they are having an idea for the first time in their life; and the person develops the solution themselves instead of learning it from someone else. However, in the case of P-creative solutions, it can be the case that the idea was developed independently by someone else before. By contrast, when a person comes up with a so-called "H-creative" solution, their idea is novel to the whole world; it appears for the first time in human history. Eminent creative figures like Mozart, van Gogh, and Einstein made contributions that were novel on a global scale by the time of their release.

The second creativity requirement, effectiveness, means that the solution functions well; it is useful, appropriate, valuable.

In addition to evaluating solutions, there is also great interest in measuring the creative capacities of individuals. Yet, as scholars have noted early on in the history of psychological studies, creativity is not a one-dimensional capacity. Rather, multiple sub-abilities are involved [9], [10]. Correspondingly, standard creativity tests assess several different constructs to estimate the person's overall creative capacities.

In the Alternative Uses Task (AUT), ordinary objects are named and test-takers are asked to think up as many uncommon uses as they can in a given timeframe [9]. For instance, if the ordinary object is a brick, it can serve as a building block – the ordinary use. However, it can also be used as a bookend, to drive a nail, it could be pulverized to produce powder for paint; the term "supreme" could be printed on it and then it might be sold for twice the normal price, etc. Test takers achieve a high score regarding *fluency* when they produce many different ideas in the given time. However, all ideas produced by the person could fall into a single category of solutions, for example, invoking the brick as a building material. In this sense, the test-taker might suggest using the brick to build a house, a church, a hospital, etc. Scores on the dimension of *originality* depend on the unusualness of ideas. A test-taker who comes up with an uncommon use for the object that hardly any other test-taker thought of receives a high originality score. The construct of *flexibility* refers to the number of different solution-categories a test-taker thinks of. A flexible thinker can be mindful of many different properties the object has and suggests corresponding uses. For example, they can consider the weight, volume, color, saleability, texture, heat conductivity, etc. to generate many different ideas for uncommon uses.

In other test approaches, people are asked to report on their creative achievements in fields such as music, creative writing, science, dance, visual arts, cooking, and more [11]. For instance, in the field of music people could report levels of achievement between (0) "I have no training or recognized talent in this area" to (7) "My compositions have been critiqued in a national publication."

To probe the feasibility of our game as an environment for creativity measurements, we begin with the standard definition of creativity: A solution is *creative* when it is *novel* and *effective*. In the game, we endeavor a respective measurement by looking at the diversity of event-chains that lead to a halted enemy (cf. section III-C). Each chain reflects a solution produced by the player that is effective because the chain culminates in the halting of an enemy. Moreover, each chain that is different from others ("diversity") is a novel chain. This is an intra-individual measure; it is a measure for P-

[2]https://www.mturk.com/

creativity. In addition, chain novelty could also be assessed by making comparisons to chains produced by all other players in the online game; this would be an approximation of a measure regarding H-creativity. Such a comparison across chains of different players becomes more informative once more individuals have participated in the game study and conclusions about "rarity" or "H-creativity" are supported by a great(er) amount of data.

Should it be feasible to assess creativity with the online game, our longer-term goal is to assess a variety of constructs that characterize the creative mindset. In this regard, our main theoretical basis is provided by John E. Arnold [12], a pioneer of creativity and innovation studies. His set of constructs includes fluency, originality, and flexibility as already mentioned above, but goes considerably beyond them. Other constructs of interest are problem-sensitivity (the inclination of a person to notice and tackle problems), daringness (the willingness of a person to challenge the status quo and risk the untried), drive (the emotional energy and enthusiasm with which a person pursues their project, specifically when facing hardships) and more. Ideally, at some later stage of game development, all these different constructs can be assessed validly by analyzing people's gaming behavior.

B. Creativity in Video Games

Close relationships between playing and creativity have often been noted in the literature. This is no different when it comes to video games. They provide a large possibility space for players [13]. From a well-defined starting point, there are usually many ways to reach the end state. These possibilities can free the player from straightforward thinking and allow them to find novel solutions by exploring even unrealistic or improbable options [14]. Whether it is the creation of a character in a role-playing game, finding strategies in a first-person shooter, or modifying the game world in a sandbox game, many video games require some form of creativity of their players.

Csíkszentmihályi [15] states that highly creative people often enter "the flow," a period of complete focus with a high level of experienced fulfillment. This is also what many game designers try to invoke in their players, by creating an environment with the same key components: Clear goals, no distractions, direct feedback, and continuous challenge [16]. Moreover, it has been found that certain types of video games can have a positive effect on players' creativity [17], [18].

Other projects concerning the measurement of creativity have used games or game-like tests. Hart et al. [19] propose an automated test to gain insights on creative exploration. Players create shapes by rearranging ten cubes into patterns they deem interesting or beautiful. The authors find alternating phases of exploration, where participants shift rapidly between many different types of shapes, and exploitation, where participants create many similar shapes in quick succession. This game measures fluency through the number of shapes a participant creates and originality by the overall rarity of each shape. The authors observe that these in-game measures correlate with AUT creativity scores.

III. Approach

This section provides detailed information about our design goals and how they enable us to measure creativity. A description of one example game round and the resulting creativity measurements is provided.

A. Design Goals

Our approach to a game design that requires creativity from its players and makes it possible to measure creativity is led by several key design goals.

To allow for exploration, there should be a high number of options for the player to choose from. These can either be distinct options, such as building a tower or placing spikes, or combinations of upgrades that change the way towers behave. Furthermore, effects that arise from a combination of terrain features (such as water) and tower effects provide a larger creative space to explore. To encourage players to try out different strategies we put them under pressure, hoping that players will attempt to explore unconventional, unexpected strategies in an attempt to find a way to still beat the game. Pressure can arise from a shortage of resources or a difficulty level that appears too high, for example too many enemies spawning. As such, the game needs to either constrain the amount of money that players have at their disposal to buy defenses or spawn a number of enemies that cannot be fended off with a naive strategy.

Another important design goal is that the game's core mechanics need to be fully understood within a short time frame without any human supervision. This is crucial should the game run on MTurk, where direct real-time experimenter intervention is not possible. If players perform badly because they do not understand how objects can be placed, for example, their results would be useless or misleading for measuring creativity. At the minimum, these need to be identified and discarded in the data analysis.

To ensure that players can predict and comprehend effects within the game, we designed it around typical simplifications of real-life physics: water conducts electricity, trees start burning when coming into contact with fire, ice slows movement, etc. To encourage exploration, there is also a set of non-obvious effects, such as stones exploding if first frozen and then hit with fire.

The design elements of the game should not distract the player, either visually or thematically, as to not skew the results. As a consequence, we decided to design a highly abstracted scenario, where players need to combat an unambiguously "evil" enemy, in this case, bacteria and viruses.[3] The abstractness of the scenario should reduce subjective experiences of "violence," to be mindful of ethical aspects in the design of games for research purposes.

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.

B. Gameplay

Each round of the game lasts for five minutes, independent of player performance. Before the game begins players are shown a tutorial that introduces the main game mechanics. The objective of the game is to protect the heart from bacteria and viruses, as depicted in figure 1. Throughout the game, multiple enemy types are introduced.

Early enemies have very low health and can be defeated easily by the player. This allows players to learn the game's functionality without added pressure. Then, more difficult enemy types start to spawn. First, bacteria appear that have high health but move slowly. Next, viruses emerge, which move more quickly and can replicate themselves when adjacent to boulders. Lastly, bacteria appear that have low health but are immune to fire. Most of the time only one type of enemy spawns to allow players to adapt their strategies. Short periods with multiple enemy types spawning also occur to allow players to express their creative flexibility. Whenever an enemy reaches the heart it is damaged. This is reflected in a red damage bar beneath the game menu, which fills up from left to right. Each time the bar is filled completely the heart appears more damaged.

[3]The immune system theme was conceived before the COVID-19 pandemic.

The player receives coins slowly over time. These can be used to purchase defenses and to place them along a grid on the play area. The main defenses are towers that shoot antibodies at the advancing enemies. Towers can be upgraded to have elemental powers (fire, ice, and electricity), as well as enhanced range and shooting speed. In addition to towers, other game elements can also be purchased by the player: boulders (stopping enemies from advancing for a while), water (making enemies wet that pass through), trees (spreading fire), and spikes (damaging enemies that pass through them). Combining different effects can lead to greater efficiency: An enemy that burns might set fire to a tree, which in turn can set more enemies on fire; freezing a wet enemy with an ice antibody can stop it and other enemies from advancing too quickly; freezing and then heating boulders in quick succession causes them to explode, inducing damage to all enemies within a certain range.

C. Game Measurements

To analyze the players' behavior, gameplay events are recorded. These include events triggered directly by the player as well as game events developing as a consequence of the player's action. Events directly triggered by the player include placing objects and upgrading towers. Other events keep track

of when enemies spawn, die or reach the heart, as well as each time an enemy or other object changes its state, for instance when it is set on fire. Each event is converted into a log entry, with timestamp and event location, and identifier and type of the object which caused the event. For events that track a change of state, it is also noted which object caused the change.

These log entries allow the construction of event chains that lead to an enemy death. Whenever an event is caused by a second entity, all previous events in that entity's chain are also included. For example, if an enemy was electrified after entering electrified water, all events that lead up to the water coming in contact with the electricity are also considered part of the chain, since without these events the enemy would not have been defeated.

The diversity of these chains is the first major measure to assess a player's creativity. It builds on the standard definition of creativity. Each chain that leads to the halting of an enemy is an effective chain. Each chain that differs from others (chain diversity) is a novel chain. This creativity score is defined by the following function, where N is the number of chains:

$$\sum_{i=0}^{N} \sum_{j=i+1}^{N} \frac{Levenshtein(chains[i], chains[j])}{(N(N-1))/2}$$

We use the Levenshtein distance, a metric to determine the difference between two strings. The Levenshtein distance is the minimum amount of single-character edits needed to transform one string into another. Edits can be character insertions, deletions, or substitutions. For example, the Levenshtein distance between "tower" and "tree" is 3. Two characters are substituted (*o* and *w*) and the character *r* is deleted.

To be able to calculate Levenshtein distances, each event type is assigned a distinct character and each event chain is transformed into a string. The chain "start burning (fire), get wet (water), become frozen (ice), death" would be transformed into the string "fwid." For each player, Levenshtein distances for all chain pairs are summed up and then averaged, as to not favor players that managed to defeat more enemies.

IV. Method

In this section, we first share some insights on how a pilot study informed the main study. We then provide demographic data from participants of the main study, review the instruments used and the study procedure.

A. Game Changes Between a Pilot Study and the Main Study

We conducted a smaller pilot study prior to the one described in this paper. Participants from the pilot study were not allowed to take part in the main study. The data gathered in the pilot was not evaluated due to the game still being experimental with a few errors. Nonetheless, we gained useful insights to refine our game, in particular regarding to its applicability in creativity research contexts.

In the game version used in the main study, the game length is fixed for all participants, independent of their performance. In the prior game version, the game could be "lost" if the player let a certain amount of enemies near their heart. The resulting data was difficult to compare, as some games were lost in a short amount of time.

There are no random elements anymore. In the pilot game version, enemy types would spawn based on a random factor, which was replaced by a uniform timer to spawn enemies. This reduces noise in the gameplay data considerably.

Also, the current game version has more elements available, such as spikes, electricity towers, new types of enemies, and the possibility for boulders to explode. This gives players more options to explore different gaming strategies.

Towers now have a limited range in which they can damage enemies. Reducing the area in which a tower shoots helps to balance the game since one tower alone does not suffice anymore to hit enemies everywhere on the game map. Having a more specific range also facilitates the development of more intricate strategies such as intentionally setting trees on fire with a nearby fire tower.

Most importantly, a short tutorial is shown before the first game round. This helps to make the game mechanics clear from the start. In the pilot study, some participants lost quickly because they did not understand basic game interactions (e.g., how to click on a tower to place it), which rendered their data unusable for evaluation purposes.

B. Participants

The main study was conducted with N=17 participants, 6 female, 11 male, all affiliated with a Digital Engineering Faculty. 14 participants fell in the age range of 18 to 27; they were bachelor and master students. Three participants fell in the age range of 28 to 37; they were Ph.D. students and postdocs.

To better characterize the sample of participants, a questionnaire inquired about previous experiences with the AUT and video games. Nine participants stated they had never participated in an AUT or similar test, the remaining eight participants had taken part at least once. Only one participant answered to never play video games. Six stated that they played daily and ten selected to play weekly or monthly. Eleven participants selected that they were familiar with Tower Defense games as a genre.

All participants received a small compensation in the form of a chocolate bar or gummy bears for taking part in the study.

C. Instruments

Three types of instruments were used to gather data.

1) Alternative Uses Test: The AUT, as described in section II-A, provides measures on several sub-scales of creativity – most prominently *fluency*, *originality* and *flexibility*. The ordinary object in our test was a paper clip. Three raters assessed test responses independently. The *fluency* rating for each participant is given by the number of ideas each participant submitted. *Originality* is defined by how rare an option is within the overall population. We rated the originality of each idea on a scale from 0 to 4. A rating of 0 means that a test response is not original at all, but describes the intended use

of the paper clip. 1 was assigned to ideas that were named by more than two other participants. 2 was given when an idea was named by one or two other participants. When an idea was only proposed by the test-taker it received a 3 when it appeared rather obvious or random, and it received a 4 when it seemed thoughtful or cunning. For the originality rating of a single idea, we determined the mode of the three expert ratings. If no mode could be found, we instead used the median. The final originality score of a participant is the maximum of all their idea scores. For the *flexibility* rating, we assigned categories to all ideas. The flexibility score of a participant is the number of different object-usage-categories that were inferred from the participants' ideas.

2) Questionnaire: Data to better characterize the sample of participants, such as people's pre-experiences with computer games, was gathered via an online questionnaire.

3) Gameplay Data: We extracted two main measures from the gameplay data: The Levenshtein score as described in III-C is a creativity metric; it assesses how many novel and effective game sequences the player produced. The game score counts how many enemies the player neutralized and is thus a measure of mere gameplay effectiveness.

D. Procedure

Participants took part in the study on computers provided by us. The study consisted of three parts. In the first part, the AUT, participants were asked to name as many uses for a paper clip as they could within three minutes. Afterward, each participant filled out an online questionnaire about their previous experiences with video games and AUT. Lastly, the participants played the game once. The game started with a tutorial and then lasted for five minutes. The tutorial demonstrated the game controls and main mechanics. It showed where enemies will spawn and how they move towards the heart, and that players will gain money over time. It showcased how to use the menu to buy buildings and upgrades. However, it did not explain the specific function of the game elements, only that towers will shoot projectiles at enemies within range. The tutorial did not explain all game elements exhaustively so that players would be encouraged to try them out during the gameplay, it was nonetheless revealed that combining different effects might be more effective than each mechanic on their own.

We tried to give the participants as little help as we could, to see if it would be viable to run a similar study on MTurk. Only in cases where a participant struggled to understand the basic game interactions, that is when they tried to use drag and drop instead of clicking, we provided them with some tips. We did not explain game mechanics, only game controls. This help was only needed at the very beginning of the game and thus should not impact player performance. Only a few of the participants needed this intervention to play the game.

V. Results

In this section, we overview study results and descriptive statistics, and then analyze patterns in the data.

A. Descriptive Statistics

Table I provides an overview of key variables in this study.

Variable	Min	$\bar{x}$	Max	s
AUT Fluency	5.0	8.0	17.0	3.1
AUT Flexibility	4.0	7.1	13.0	2.3
AUT Originality	1.0	2.9	4.0	1.0
Levenshtein Score	0.8	1.4	2.3	0.5
Game Score	71.0	117.6	178.0	28.9

TABLE I
DESCRIPTIVE STATISTICS REGARDING KEY VARIABLES IN THE MAIN STUDY (N = 17)

Measures of AUT fluency can range between zero and any positive integer, depending on how many ideas the participant wrote down in the given time-frame. In our study, the most fluent participant wrote down 17 ideas, which is several times more than the least fluent participant with 5 ideas.

Values of AUT flexibility can range between zero and the number achieved by the person on the fluency scale. If two or more ideas noted by the participant fall in the same object-usage category (e.g., two different ways in which a paper-clip can be used for body ornamentation), then the person's flexibility score is lower than their fluency score. In the present study, the participants' flexibility scores ranged between 4 and 13, so that the highest-scoring participant scored several times higher than the least scoring person.

AUT originality values can range between 0 and 4 in this study. Least scoring participants received a value of 1; they came up with ideas only that were relatively common in the overall group (all their ideas were also proposed by at least two other study participants). Highest-scoring individuals achieved the maximum value of 4 on the originality scale.

Thus, overall, a considerable variance is found on all three creativity-dimensions, which is a traditional and typical finding in creativity research [20].

Levenshtein scores are average measures based on the person's Levenshtein chains (and how much the chains differ from one another). The longest chain produced by an individual in our study consisted of 10 game-events, but the average chain length was 2.5. Thus, Levenshtein scores are confined between zero (when the person doesn't halt an enemy or only produces the same event-chains again and again) and the maximal chain length produced by the person. In this study, Levenshtein scores were found between 0.8 and 2.3.

Game scores can range between zero and the number of enemies spawned in the game, 221 in this study. Once again, participants with the highest values on this dimension obtained scores larger by orders of magnitude than study participants with the lowest values.

B. Patterns in the Data

Concerning individual analyses, it can be noted that the two highest-scoring participants on the flexibility dimension also produced the highest Levenshtein scores.

Concerning the overall data-set, an overview of how different measures relate to each other is provided in table II.

		1	2	3	4	5	6
1.	AUT Flexibility	1					
2.	AUT Fluency	0.95*	1				
3.	AUT Originality	0.14	0.20	1			
4.	Levenshtein Score	0.57*	0.54*	0.13	1		
5.	Game Score	0.36	0.34	0.12	0.06	1	

TABLE II
THE TABLE SHOWS PEARSON CORRELATIONS, EXCEPT FOR AUT ORIGINALITY, WHICH IS CACULATED WITH KENDALL'S TAU. * $< .05$ SIGNIFICANCE

While fluency is a sine qua non for creativity (in particular, a person can only produce so many diverse ideas as they produce numbers of ideas in total), flexibility is of greater theoretical interest as it measures directly the diversity of a person's ideas. Therefore, table II presents flexibility on top to ease the visual comparison with all other measures.

Notably, the Levenshtein score correlates significantly with AUT flexibility ($r = 0.57$, $p < 0.05$).

Since AUT flexibility scores depend to a considerable extent on AUT fluency scores, a significant correlation between the two is a typical finding and applies to this study as well. Consequently, Levenshtein scores also correlate with AUT fluency ($r = 0.54$, $p < 0.05$).

Further statistically significant correlations are not found in this study with N=17 participants. Yet it can be noted that, consistently, all three AUT creativity dimensions relate positively to Levenshtein scores.

One key question concerning the usefulness of this game and the Levenshtein score as a means to measure creativity concerns the impact of game pre-experiences. It could be the case that people who play regularly perform better in Immune Defense, and they might perform more complex and diverse game actions, simply because they know how to play computer games. If such relations existed, high Levenshtein scores might indicate (a) that the participant is creative, (b) that the participant has a lot of gameplay experience, or (c) that both is the case. Therefore, it is important to clarify how Levenshtein scores relate to pre-experience in gameplay. Table III provides an overview.

		1	2	3
1.	Levenshtein Score	1		
2.	Game Score (Performance)	0.07	1	
3.	Gameplay Pre-Experience	-0.19	0.00	1

TABLE III
THE TABLE SHOWS CORRELATIONS BY KENDALL'S TAU DUE TO THE ORDINAL SCALE LEVEL OF MEASURES ON GAME PRE-EXPERIENCE

The Levenshtein score is not related to game performance in this study (cf. VI). Moreover, there is no statistically significant relationship to gameplay pre-experience. If non-significant tendencies were to be interpreted, experienced gamers rather appear at a slight disadvantage: Their Levenshtein scores tend to be a bit lower than scores of inexperienced players (r=-0.19, p=.37). It can be added that the (surprising) zero-correlation between game scores and pre-experience in gameplay is not a typo, but a numerical result we checked several times; it will also be re-considered in the discussion.

To clarify the underlying structure of the variables, we calculated a factor analysis (Principal Component Analysis). This approach reduces the dimensionality of data. It searches for a low(er) number of "underlying factors" to explain the variance in a given data set. In our study, the PCA yielded a two-factor solution, cf. table IV.

		Factor 1 "Creativity"	Factor 2 "Regular Gameplay"
1.	AUT Fluency	.96	.06
2.	AUT Flexibility	.93	.18
3.	AUT Originality	.44	-.54
4.	Levenshtein Score	.65	.20
5.	Game Score	.39	.60
6.	Gameplay Pre-Experiences	-.53	.65

TABLE IV
LOADINGS OF ORIGINAL VARIABLES ON TWO UNDERLYING FACTORS FOUND WITH A PRINCIPAL COMPONENT ANALYSIS

Methodologically, study methods are known to create artifacts, so that all data points generated by one method A tend to cluster together, whereas all data points gathered by another method B by tendency also cluster together [21]. In our study, we used three different methods to acquire data, the AUT, the introductory questionnaire, and the video game. If method artifacts were strong in our study, a factor analysis would find one or more "method factor(s)" as explaining a considerable amount of data variation. In this case, an underlying factor might be found where all AUT-based measures load highly (correlate positively and significantly), whereas all data points gathered with other methods would not load highly (correlate not at all or correlate negatively).

Because method artifacts can be strong it is noteworthy that this PCA does not find such a method factor. Rather it produces a two-factorial solution that seems to distinguish starkly between "creativity" (factor 1) versus "regular gameplay" (factor 2). All AUT creativity-indices and the Levenshtein score load highly on factor 1. By contrast, game scores and questionnaire-based information on pre-experience with video games load highly on factor 2.

VI. DISCUSSION, LIMITATIONS AND FUTURE WORK

This study explores a novel approach for the measurement of creativity, which promises to bring about benefits such as greater automation of creativity assessments, studies with much-increased numbers of participants, fewer possibilities for participants to manipulate their scores compared to creativity self-ratings, and more ecological validity of the procedure compared to the artificial tasks used in many standard creativity tests.

The key question in this study concerns feasibility: Can we measure creativity with an online game? Or does people's gaming behaviour only reflect other variables, such as people's gameplay pre-experience, rather than creativity? The study results show that it is indeed possible to measure creativity with an online game like Immune Defense. Moreover, the first-probed metric of Levenshtein scores shows great promise as a means to capture creativity and not game performance or

gameplay pre-experience. Levenshtein scores of participants correlate significantly with people's scores in a standard creativity test (the AUT). Levenshtein scores do not correlate with people's game scores or gameplay pre-experience.

Moreover, a factor analysis across our key variables suggest that the Levenshtein score clusters with AUT creativity measures, not with regular gameplay variables. That means, even though the Levenshtein score is calculated based on observations in the online game, it is more related to what people do and achieve in the AUT creativity assessment than to people's gaming habits. This is a very promising outcome, which encourages much further probing of more creativity-indices that can be derived from observations of gameplay behaviors.

Based on these auspicious findings, a considerable number of further indices can be explored in the future, and they should be tested in a much larger and more variegated sample of participants. In this context, we can also re-consider the surprising zero-correlation between people's reported pre-experience with computer games and their Immune Defense scores. In this study, all participants were affiliated with a Digital Engineering faculty, where people are familiar with computers in general and usually also with computer games. While the participants' self-reported gameplay pre-experience varied somewhat, in a sample reflecting the overall society there clearly would be even greater variation. So, in a more heterogeneous sample of participants, a positive correlation between Immune Defense game scores and gameplay pre-experience would most probably be found.

Since it seems feasible to assess creativity with the Immune Defense game, we can now proceed to realize a greater bandwidth of design objectives. In particular, more metrics like the Levenshtein score need to be tested to cover the whole suite of creativity dimensions differentiated in section II. For subsequent validation studies, it will also be important to include more standard creativity tests beyond the AUT. While the AUT provides values for most common creativity dimensions like fluency, originality, and flexibility, it does not cover all dimensions of interest. For instance, the AUT does not deliver measures on the creativity dimensions of problem-sensitivity, drive, and daringness.

Thus, follow-up studies need to size up in several regards: more study participants, more heterogeneity in the sample, more standard creativity tests used for validation purposes, and different creativity scores derived from people's gameplay behaviors beyond the Levenshtein score.

For now, we can say that it is possible to measure people's overall creativity with the Immune Defense game and the Levenshtein metric. In the near future, hopefully, we will be able to offer more fine-grained analyses of people's creative capacities after just five minutes of gameplay observations. In the not-so-near future, it might be possible to re-use metrics developed for Immune Defense and automate creativity assessments in other contexts. Maybe it becomes possible to assess a range of creative capacities by observing people's behaviors in games other than Immune Defense, in the way people write e-mails, how they behave in video calls, or how they act in yet other contexts. Automated creativity assessments are possible.

REFERENCES

[1] K. Paul, E. S., *The Philosophy of Creativity: New Essays*. Oxford University Press, 2014.

[2] M. Kieran, "Creativity as a virtue of character," in *The Philosophy of Creativity*. Oxford University Press, May 2014, pp. 125–144. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199836963.003.0007

[3] C. Zhou, Ed., *Handbook of Research on Creative Problem-Solving Skill Development in Higher Education*. IGI Global, 2017. [Online]. Available: https://doi.org/10.4018/978-1-5225-0643-0

[4] G. Hawthorne, M. Saggar, E.-M. Quintin, N. Bott, E. Keinitz, N. Liu, Y.-H. Chien, D. Hong, A. Royalty, and A. Reiss, *Designing a Creativity Assessment Tool for the Twenty-First Century: Preliminary Results and Insights from Developing a Design-Thinking Based Assessment of Creative Capacity*. Springer International Publishing, 05 2016.

[5] M. A. Runco and G. J. Jaeger, "The standard definition of creativity," *Creativity Research Journal*, vol. 24, no. 1, pp. 92–96, Jan. 2012. [Online]. Available: https://doi.org/10.1080/10400419.2012.650092

[6] M. Boden, "The creative mind: Myths and mechanisms: Second edition," *The Creative Mind: Myths and Mechanisms: Second Edition*, pp. 1–344, 01 2003.

[7] J. Kaufman and R. Beghetto, "Beyond big and little: The four c model of creativity," *Review of General Psychology*, vol. 13, 03 2009.

[8] O. Kolodny, S. Edelman, and A. Lotem, "Evolved to adapt: A computational approach to animal innovation and creativity," *Current Zoology*, vol. 61, no. 2, pp. 350–368, 2015.

[9] J. P. Guilford, "Creativity." *American Psychologist*, vol. 5, no. 9, pp. 444–454, 1950. [Online]. Available: https://doi.org/10.1037/h0063487

[10] E. P. Torrance, *Torrance Tests of Creative Thinking: Norms–Technical Manual Research Edition—Verbal Tests, Forms A and B; Figural Tests, Forms A and B*. Princeton, NJ: Personnel Press, 1974.

[11] S. Carson, J. Peterson, and D. Higgins, "Reliability, validity, and factor structure of the creative achievement questionnaire," *Creativity Research Journal*, vol. 17, pp. 37–50, 12 2005.

[12] J. von Thienen, W. Clancey, G. Corazza, and C. Meinel, *Theoretical Foundations of Design Thinking. Part I: John E. Arnold's Creative Thinking Theories*. Springer International Publishing, 01 2017, pp. 13–40.

[13] W. Wright. (2006). [Online]. Available: https://web.archive.org/web/20200321122928/https://www.wired.com/2006/04/wright-2/

[14] C. Mainemelis and S. Ronson, "Ideas are born in fields of play: Towards a theory of play and creativity in organizational settings," *Research in Organizational Behavior*, vol. 27, pp. 81–131, Jan. 2006. [Online]. Available: https://doi.org/10.1016/s0191-3085(06)27003-5

[15] M. Csíkszentmihályi, *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, 1996.

[16] J. Schell, *The Art of Game Design: A Book of Lenses*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[17] J. A. Blanco-Herrera, D. A. Gentile, and J. N. Rokkum, "Video games can increase creativity, but with caveats," *Creativity Research Journal*, vol. 31, no. 2, pp. 119–131, Apr. 2019. [Online]. Available: https://doi.org/10.1080/10400419.2019.1594524

[18] L. A. Jackson and A. I. Games, "Video games and creativity," in *Video Games and Creativity*. Elsevier, 2015, pp. 3–38. [Online]. Available: https://doi.org/10.1016/b978-0-12-801462-2.00001-1

[19] Y. Hart, A. E. Mayo, R. Mayo, L. Rozenkrantz, A. Tendler, U. Alon, and L. Noy, "Creative foraging: An experimental paradigm for studying exploration and discovery," *PLOS ONE*, vol. 12, no. 8, p. e0182133, Aug. 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0182133

[20] F. Barron, "The disposition toward originality," *Journal of Abnormal and Social Psychology*, vol. 51, no. 3, p. 478–485, 1955.

[21] D. W. Campbell, D. T. & Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix," *Psychological Bulletin*, vol. 56, pp. 81–105, 1959.