Detecting who is going to innovate

DR KATHARINA LOCHNER (corresponding author)
University of Applied Sciences Europe, Germany
Katharina.Lochner@gmail.com

DR ACHIM PREUSS
cut-e an Aon company, Germany
Achim.Preuss@cut-e.com

RICHARD JUSTENHOVEN
cut-e an Aon company, Germany
Richard.Justenhoven@cut-e.com

The authors

Dr Katharina Lochner is a researcher and lecturer in the areas of Industrial and Organisational and Positive Psychology at the University of Applied Sciences Europe in Iserlohn, Germany. In this role, based on her more than twelve years of experience in the area of psychological assessment working with different clients from various industries, she connects research and practice of work and organisational psychology.

Dr Achim Preuss is a renowned pioneer in the assessment industry and a visionary practitioner. As Head of Global Solutions at Aon's Assessment Solutions, he is responsible for the company's global product development and its best practice innovations in preapplication attraction and selection. His main areas of expertise are job analysis, knowledge engineering, and artificial intelligence.

Richard Justenhoven is a leading organisational psychologist and an acknowledged expert in the design, implementation and evaluation of online assessments.
As Product Development Director at Aon's Assessment Solutions he oversees the creation of groundbreaking off-the-shelf products and bespoke client solutions that stem from technological innovations and the application of business psychology.

Abstract

An important factor for workplace innovation is having creative employees in the workplace. Most tests for assessing creativity are paper-and-pencil-based and require a trained evaluator for scoring. This, in turn, renders scoring time-consuming, expensive, and not entirely objective. The aim of the three studies presented here was to develop an online creativity test that uses a fully automated scoring algorithm, that is optimised for unsupervised settings, and that can be applied internationally by being language-independent. As such, it can be used as a quick and cost-efficient instrument for selection as well as individual and team development purposes.

Introduction

Innovation has become something like a "Holy Grail" for organisations, given that innovative products and services provide a competitive advantage in rapidly changing international markets (Maier, Streicher, Jonas, & Frey, 2007). Moreover, organisations see workplace innovation (WPI) as a factor to help them face the challenge of today's volatile markets (Kesselring, Blasy, & Scopetta, 2014). Thus, companies, on the one hand, strive to establish an environment that facilitates innovation (Amabile, Conti, Coon, Lazenby, & Herron, 1996) and, on the other hand, try to recruit innovators (i.e., people that are likely to innovate).

But how can one assess during the selection process who will most likely be an innovator? Whether someone will be an innovator is determined by a number of factors, such as cognitive ability (i.e., intelligence), certain personality characteristics, such as openness to experience, and creativity (Farr, Sin, & Tesluk, 2003; Soosay, 2005; Streicher, Maier, Frey, Jonas, & Kerschreiter, 2006). All of these characteristics can be measured by using psychometric tests and questionnaires.

There are different modes of administering such tests and questionnaires: either in paper-and-pencil mode or online (online assessment). Paper-and-pencil mode usually implies that participants are invited for an on-site session during which they fill in the instruments on paper with an administrator being present. For this mode, there are a variety of instruments available that assess the above-mentioned personality characteristics, cognitive abilities, and creativity.

During the past few years, however, online assessment (i.e., tests and questionnaires adiminstered via the internet) is becoming more and more popular with recruiting companies: In a 2012 survey of European companies, 83% of them indicated using online assessment in their recruitment processes (cut-e, 2012). Moreover, a recent article in the Harvard Business Review (Bateson, Wirtz, Burke, & Vaughan, 2013) recommends using unsupervised online assessment as a first stage in the process of employee recruitment.

Thus, it would be desirable to be able to measure applicants' potential to innovate in unsupervised online mode. This mode requires candidates to be able to complete the assessment using a computer and the internet and it usually involves automated scoring and reporting.

There are instruments for measuring cognitive ability and personality in unsupervised online settings using automated scoring. However, to date, creativity tests have mostly required supervised settings and always a trained evaluator. Training evaluators takes time and, even if there are two or more evaluators per test, it will still mean that there is a certain amount of subjectivity involved. Moreover, evaluation usually takes about ten minutes per test, so evaluating 100 tests will take over 16 hours, assuming there is only one evaluator per test. This is time-consuming and expensive.

Thus, the aim of the three studies presented here was to develop an online creativity test that uses a fully automated scoring algorithm and that is optimised for unsupervised settings. To this end, Study 1 is an exploratory study designed to find and combine parameters that can be used for an automated scoring algorithm. Study 2 evaluates the first part of a scoring algorithm designed on the basis of Study 1 by comparing its results to the rating done by trained evaluators. Finally, Study 3 validates the full scoring algorithm as well as the creativity test designed based on the insights from Studies 1 and 2.

Assessing creativity

Creativity is often equated to divergent thinking. Divergent thinking (Guilford, 1950) generates creative ideas by exploring various possibilities. In contrast to convergent thinking (i.e., a process that focuses on coming up with a single, well-established answer to a problem) in which a series of logical steps are diligently connected in order to reach a goal, divergent thinking (or lateral thinking; De Bono, 1992) is spontaneous. Many possibilities are gone through in as little time as possible and with the least possible evaluation, thereby giving rise to new connections between elements. Options are produced by decomposing the problem into conceptual parts and then recombining them.

In the literature, creativity is often seen as consisting of four different components, fluency, flexibility, and originality (e.g., Jäger, Süß, & Beauducel, 1997) as well as elaboration (Torrance, 1966). Fluency is the number of unique responses within a certain period of time (Guilford, 1967; Torrance, 1966). Flexibility refers to the extent to which the responses relate to a diverse range of categories and perspectives and the use of broad and inclusive cognitive categories (Guilford, 1967; Torrance, 1966). Originality refers to the extent to which responses are novel, infrequent and uncoventional (Guilford, 1967; Torrance, 1966). Elaboration is the amount of detail given in each response (Guilford, 1967, Torrance, 1966).

A well-established and validated instrument for assessing creativity in a supervised setting with trained evaluators is the Torrance Tests of Creative Thinking (TTCT; Torrance, 1974). This instrument is directly based on Guilford's (1950) work. TTCT have the most extensive empirical basis of all measures of creativity so far, and longitudinal studies prove the effectiveness of the underlying concept (Torrance, 1981). In a TTCT the participant has to use a number of given shapes, combine or complete them, and name the result.

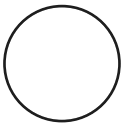
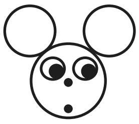
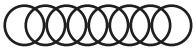
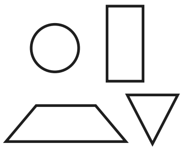
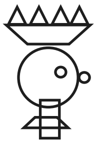
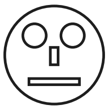
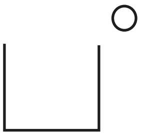
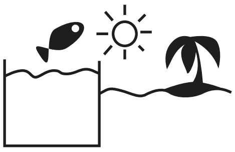
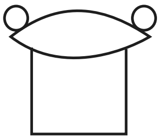
Figure 1 depicts the different types of tasks a TTCT uses and provides examples of more or less creative solutions for the tasks. When comparing the more and less creative solutions for all three types of tasks a few things are worth noting. The more creative solution for the 'Use' task involves using the shape in different sizes. The more creative solution for the 'Complete' task involves using a greater number of shapes in different sizes and orientations. Finally, the more creative 'Complete' task consists of more elements and the title is more complex and unusual.

The test is a paper-and-pencil test, meaning that the participant has to draw the objects. The results are evaluated by previously trained evaluators. In the original version, the raters scored participants on Fluency (number of objects), Flexibility (number of different categories of objects), Originality (number of unique or rare objects), and Elaboration (amount of detail in the objects). In the third edition of the TCCT dating from 1984, "Resistance to Premature Closure" (i.e., going beyond the logical way to complete a figure) and "Abstractness of Titles" (i.e., level of abstractness of the titles given to the objects) were added as rating categories.

Participants have to draw the objects manually. This means that their results are, to a certain extent, dependent on their ability to draw. Moreover, the test requires being evaluated by trained evaluators, making the process time-consuming and not one hundred percent objective. Finally, it is not suitable for unsupervised online administration.

As already mentioned, the purpose of the three studies reported here was to develop a creativity test that that can be administered in an unsupervised online setting and that can be evaluated automatically. Since TTCT is a well-established and validated test, we decided to use this test as a basis for the online test to be developed.

Figure 1. Different types of tasks used in TTCT.

	Starting Shapes	**Completed Drawing**	
		More Creative	Less Creative
Use		 Mickey Mouse	 Chain
Combine		 King	 Face
Complete		 A fish on vacation	 Pot

Note: Retrieved from https://kathrynwelds.com/2015/07/08/greater-hemispheric-specialization-not-integration-increased-creativity/

In the following sections, the three studies that were conducted to design and validate the test will be described.

Study 1

Study 1 was a pilot study designed to explore whether it is possible to find parameters in the pictures and names that result from an administration of a TTCT type test that can be used to design a scoring algorithm. To this end, a version of the TTCT was administered in paper-and-pencil mode as part of a supervised study. The purpose of the study was to find properties of the drawings and names that determine whether a solution is more or less creative.

Method

Participants

The study was conducted with *N* = 68 university students (44% female, 56% male) taking part in a competition to find the smartest student in 2014 in Frankfurt, Germany. Participants were undergraduate and graduate students enrolled for various subjects such as economics, engineering, IT, natural sciences, or social sciences at different universities in Germany and Austria. For reasons of data protection, no other biographical data apart from gender was collected.

Procedure

The competition that Study 1 was part of had been launched by a major German recruitment agency, and the prize was, in addition to winning the award as the ‘Smartest Student’, a cash prize. In order to participate in the competition, they could sign up on a website, and participation was voluntary. After registering, they completed a series of online tests and questionnaires assessing cognitive abilities (numerical, verbal, and abstract-logical reasoning as well as handling of information). Sixty-eight students with the highest test scores were invited to a one-day assessment centre-type challenge in Frankfurt. During the course of the day, participants were asked to complete the ‘Objekt-Gestaltung’ (OJ; ‘Object-design’) from the Berlin Intelligence Structure Test (BIS; Jäger, Süß, & Beauducel, 1997) paper-and-pencil test under supervision. After the instruction sequence, candidates had 15 minutes to complete as many pictures as they could within the time given.

There was no ethics commission involved for giving ethical approval for the study. However, participants were told that their participation in this study was voluntary, that it would be anonymous, and that they could withdraw from the study at any time. Afterwards, all tests were scored by two trained experts according to the guidelines given in the TTCT test manual (Torrance, 1975). The same two experts then conducted a qualitative analysis and looked at parameters that could be used to automate the scoring.

Measures

As mentioned above, the test used was ‘Objekt-Gestaltung’ (OJ; ‘Object-design’) from the Berlin Intelligence Structure Test (BIS; Jäger, Süß, & Beauducel, 1997). This is a task that originates directly from TTCT (Jäger, Süß, & Beauducel, 1997). The task is to combine circles, rectangles, triangles, and trapezoids into real world objects by drawing them and to name these objects. Participants were free in combining these geometrical figures, using as many or as few of them as they wanted. The only requirement was to use at least one geometrical figure.

The test has to be evaluated by trained evaluators based on guidelines given in the manual. The scores resulting from the test are Fluency (number of options produced), Flexibility (number of different options produced, assessed by categorising the options produced into pre-defined categories and counting the number of different categories present), and Originality (number of unique options produced, assessed by counting the number of options that do not fit into one of the pre-defined categories).

Based on the ideas they had from the BIS (Jäger, Süß, & Beauducel, 1997) and TTCT manuals (Torrance, 1975) (see above, number of shapes used, number of different shapes used, changes in size and orientation, complexity of name), two experts recorded for each picture the following properties:

- **Fluency:** Number of real-world objects drawn and named.
- **Flexibility:** This measure usually refers to the number of different options produced. Based on the reasoning that different options can show in how flexibly the available shapes are used, the following was counted: (1) number of circles, rectangles, triangles, and trapezoids used (this was counted for each shape, so for example in one picture there could be 1 circle, 1 rectangle, 0 triangles, 3 trapezoids); (2) number of circles, rectangles, triangles, and trapezoids changed in size

and/or orientation (this was counted per picture, so if there was a variation of size and/or shape in at least one of the shapes in the picture this picture was given a 1 on this score, if not it received a 0).

- **Originality:** This measure usually refers to the uniqueness of the objects produced. Based on the reasoning that uniqueness can be seen in how the shapes are combined and how frequent or unique the given names are, the following was counted: (1) number of different patterns, i.e., number of different combinations of the shapes used across different drawings; (2) frequency of names given (how frequently the same name appeared across the entire sample, e.g., "house" or "sun").

The two experts divided the pictures to be reviewed between the two of them and did the recordings according to the predefined guidelines described above. After completing their respective share of the pictures, they swapped and reviewed what the respective other expert had recorded. In the event of disagreement, they discussed until they reached agreement.

Results

Fluency.

Candidates had drawn between 6 and 36 pictures (*M* = 17.9, *SD* = 6.17).

Flexibility.

The number of circles, rectangles, triangles, and trapezoids that were used per picture ranged from 0 to 48 (some candidates just used one type of shape in a picture, that is why the number of other shapes used in this picture can be 0). Variations (changes in size and/or orientation) of these shapes within a picture ranged from 0 to 1, with 0 meaning the shape was not changed in size or orientation, and 1 meaning the shape was changed in size and/or orientation.

Originality.

There were almost 300 different patterns that candidates had used. The frequency of names ranged from 1 (e.g., "Mexicans from above") to 33 (e.g., "house"). Moreover, there were titles that consisted of one word only, but also titles that consisted of several words, and different lengths of words.

Discussion

The purpose of this pilot study had been to see whether it was possible to find rules for an algorithm to score creativity tests. These rules were meant to crystallise the expert knowledge based on the scoring rules from the manual. Hence, the rules for the algorithm were on the one hand supposed to reflect the scoring rules and on the other hand to be able to create enough variation in scores so that, later on, differentiation between candidates would be possible. Based on the results described above, it was obvious that the highest differentiation was to be achieved using the number of shapes used, the changes in size and/or orientation and the number of different patterns used. Also, the complexity and unusualness of the name given was a potential differentiator.

Based on this reasoning the following rules for the algorithm were defined:

- **Fluency**: Number of real-world objects drawn and named.
- **Flexibility**: As described above, flexibility was assessed using the number of shapes and the variations in size and/or orientation. Therefore, the reasoning was that these properties could be combined into patterns for each image and that based on these properties the algorithm could compare each picture a participant had drawn to the previous ones and assess whether the pattern was different from the others. Thus, flexibility for the algorithm was defined as the number of different patterns.
- **Originality**: As described above, this score focused on the uniqueness and complexity of the pictures drawn and the uniqueness and complexity of the titles given. The uniqueness of the picture was defined as the uniqueness of the combinations of shapes. E.g., combining a circle and a triangle in one picture and combining a square and a rectangle in the next would yield a flexibility score of 1 (since it is two different shapes that are being used), but a uniqueness score of 0 (since both pictures use two shapes without changing them in size and/or orientation). The complexity of the picture was defined as the number and diversity of objects used and the degree of variation of the objects in terms of size, position and rotation. The uniqueness of the title was defined by how unique or rare the title was. The complexity of the title was defined by the number of words used and by the length of each word.

The next step now was to validate the algorithm. This was done using two follow-up studies. Study 2 focused on validating the Originality score since this was the more complex one, Study 3 on validating the Fluency, Flexibility, and Originality scores, along with validating the entire test designed based on Study 1.

Study 2

The purpose of Study 2 was to validate a scoring algorithm for Originality that had been set up based on the insights from Study 1. For validation, a newly programmed web-based version of the test (called "sparks") was used and the Originality score generated by the algorithm was compared to ratings by trained experts. The study was conducted in a supervised setting in collaboration with a university in Hamburg, Germany.

Method

Participants

Five university students participating in a research project at their university in Hamburg, Germany, had recruited family members, friends, or fellow students for the study. The *N* = 65 participants were between 15 and 60 years old and had various educational backgrounds, ranging from high school to university degree. Due to reasons of anonymity no further biographical data is available.

Procedure

For this study, a new online creativity test, called "sparks" was developed, based on the same principles as 'Objekt-Gestaltung' (OJ; 'Object-design') from the Berlin Intelligence Structure Test (BIS; Jäger, Süß, & Beauducel, 1997). It is described in more detail in the Measures section.

As mentioned above, five university students participating in a research project recruited participants who were invited to a session in which they were in a room with a supervisor (one of the five students). Participants completed the online creativity test sparks while the supervisor was sitting behind them. Whenever they had finished one picture they let the supervisor know. Then the supervisor took a screen shot of the respective drawing and saved it into a file, and the participant went on with the next item.

This was done so that after test completion all the pictures would be available for manual scoring by trained evaluators. The procedure was repeated until time was up (15 minutes). At the same time the scoring algorithm automatically scored participants' responses. The procedure meant that there was an interruption after each object drawn. Thus, it was to be expected that candidates would complete less pictures than they would without the interruption.

As mentioned above participation in the study was anonymous. In order to be able to assign the screen shots taken during the session to the automated score for each participant, the administrators generated a code for each participant, consisting of the initials of the administrator's name and a number that was later on used to match the results in the data base with the screen shots taken during the administration.

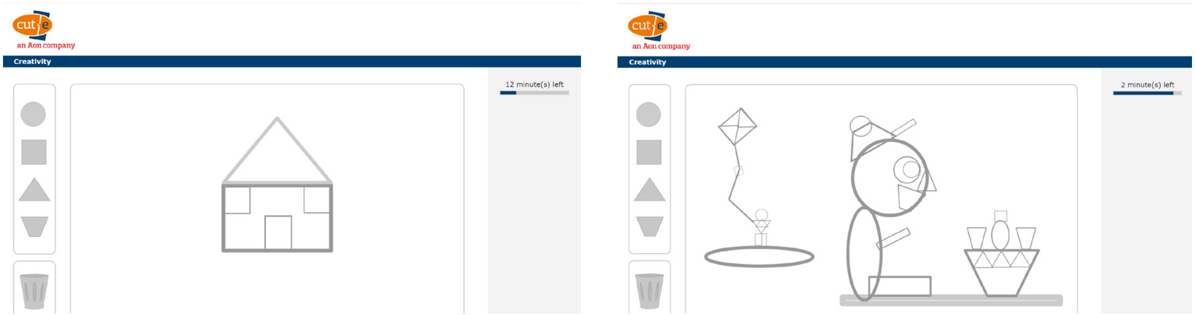
Like in Study 1, there was no ethics commission involved for giving ethical approval for the study. However, participants were told that their participation in this study was voluntary, that it would be anonymous, and that they could withdraw from the study at any time.

The five students who had supervised the test administrations were trained to evaluate the test and thus became Expert raters. Each of them received the screen shots of all participants in the study. Each of the Experts then used one excel sheet to rate all objects individually with respect to Originality since this was the most complex one of the scores. A full validation of all scores was already being planned at this stage and thus not considered to be necessary. All five Experts rated all the objects participants had drawn without knowing the other Experts' ratings. Moreover, they did not know the results of the automated scoring when evaluating the drawings.

Measures

The creativity test "sparks" that had been programmed for the study works as follows: there is a drawing area onto which participants can drag four different shapes: square, circle, triangle, and trapezoid. One shape is already present on the drawing area and has to be incorporated into the drawing, and at least one shape needs to be added. Participants can use as many of these shapes for each of their drawings as they want and they can change their size and dimensions. Once they have completed their drawing they need to name it in a text box below the drawing area. Figure 2 depicts two sample items from the test.

Figure 2. Sample items, with a less creative one on the left-hand side and a more creative one on the right-hand side



The rating criteria the Experts used were to reflect the logic the algorithm used so that it would be possible to see whether or not the algorithm had really "learned" what it had been programmed on. Thus, Originality was rated based on the following criteria: A drawing and given name was original if the participant:

- used two or more of the available shapes (e.g., at least a circle and a rectangle)
- used them in different sizes or dimensions
- combined them in a way in which they had not done it before or considerably changed the combination compared to something they had drawn before (e.g., used a triangle, rectangle, and circle once to depict a house with the sun next to it and once to depict a rocket flying to Pluto)
- added something that was unique in the category the object was assigned to (e.g., the participant combined rectangles, circles, and triangles into a forest (not unique), but added a few circles and rectangles to symbolise two people, and called it "bliss of love in the forest").

Based on these criteria the administrators assigned between 0 and 100 points to each drawing. The algorithm used the same range. Experts' ratings were averaged and then correlated with the results the automated scoring had yielded. Before doing so, interrater reliability, i.e., agreement among raters, was tested using the intra-class correlation coefficient. As a model, two-way random was chosen since the same raters rated all participants, and the type used was absolute agreement. The ICC was .95, *p*<.01, which can, according to Cichetti (1994), be interpreted as excellent. Thus, an overall rating of the scores for originality was calculated as an average of the five Experts' ratings.

Results

The results obtained by the automated scoring algorithm were compared to those the trained Experts had given.

Table 1 depicts descriptive statistics for the Originality scores each of the Experts (trained evaluators) had given and for the Originality score calculated by the automated scoring algorithm.

Table 2 depicts correlations between the five experts (individually and averaged) and the scoring algorithm. It can be seen that, based on Cohen's (1992) taxonomy, the correlations between the Expert ratings and the scoring algorithm can be classified as high.

Table 1
Descriptive Statistics for Originality as Calculated by the Scoring Algorithm, and as Rated by Trained Evaluators

	M	*SD*
Scoring algorithm	25.23	25.51
Expert 1	24.04	17.25
Expert 2	24.29	17.38
Expert 3	31.32	17.95
Expert 4	27.89	17.66
Expert 5	30.53	20.04

Note. *N* = 68. Possible range of the score = 0-100.

Table 2
Intercorrelations Between Originality Scores Calculated by the Scoring Algorithm and Rated by the Experts

	(0)	(1)	(2)	(3)	(4)	(5)	(6)
(0) Scoring algorithm	1	.78**	.72**	.57**	.73**	.73**	.77**
(1) Expert 1		1	.92**	.74**	.88**	.84**	.95**
(2) Expert 2			1	.74**	.86**	.78**	.93**
(3) Expert 3				1	.83**	.69**	.87**
(4) Expert 4					1	.82**	.95**
(5) Expert 5						1	.90**
(6) Experts averaged							1

Note. *N* = 68. ** *p* < .01.

Discussion

Intercorrelations between the five Expert ratings indicate that they followed the guidelines. The high intraclass correlation supports this. Thus, interrater reliability can be considered to be excellent, which is to be expected because they rated the drawings based on standardised guidelines from the test manual. However, what is remarkable is that the correlations between the individual and the averaged Expert ratings on the one hand and the score based on the scoring algorithm are rather high. Thus, the parameters for the scoring algorithm that were defined in Study 1 reflect an Expert rating on Originality.

One limitation of this study is that the sample was a convenience sample since the students conducting it had recruited family members and friends. Thus, the question is to what extent the sample can be seen as representative of those who are likely to take the test in a true selection setting. Moreover, in such a sample participants' motivation can be expected to be different from the motivation in a more "neutral" sample.

Another limitation of the study is that participants were interrupted during test administration because the screen shots had to be taken and saved. Thus, it is likely that they drew less pictures than they would have drawn without being interrupted. This, however, is not so much of a problem since the score we focused on in this study was Originality (not Fluency, so the number of pictures drawn). However, it is also possible that the interruptions disrupted participants' concentration and flow. Thus, it could be that their scores on Originality were lower than they would have been had they completed the test without interruptions. However, their true Originality score was not so much the focus here, rather it was the scoring algorithm, and it is unlikely that lessend concentration and focus had an impact on this.

Finally, the study had focused on validating the Originality score only, not the Fluency and Flexibility scores. Thus, an additional study was required to validate the other two scores, along with a construct validation of the newly programmed test. This was the purpose of Study 3.

Study 3

In Study 3, we focused on all three scores (Fluency, Flexibility, and Originality) as well as on the entire creativity test with respect to test-retest reliability, convergent and discriminant validity. Moreover, it was necessary to find out whether the test result depended on the ability to draw because, if so, there would be error variance in the overall score because it would assess another construct along with creativity.

Method

Participants

There were *N* = 470 participants in the study. Of those 30% were female and 70% were male. They were between 17 and 67 years old (*M* = 34.00, *SD* = 11.00) and had predominantly higher education. They all lived in the USA and completed the instrument in English.

In the re-test sample that completed sparks for a second time two weeks after the first test administration there were *N* = 120 participants. Of those 34% were female and 66% were male. They were between 19 and 57 years old (*M* = 33.00, *SD* = 10.50). Educational background was predominantly higher education. Again, they all lived in the US and completed the instrument in English.

Procedure

The study was carried out using Amazon Mechanical Turk (mTurk) in February 2015. mTurk is a platform where people are being paid for completing various kinds of tasks online in front of their home computer, for example for completing tests and questionnaires that are part of a study. Online studies are becoming more and more popular, which raises questions regarding the reliability and interpretability of the results. Studies indicate that participants in these kinds of research designs are motivated and diverse (Lochner, 2016). McGraw, Tew, and Williams (2000) collected data from online experiments across a period of 2 years and came to the conclusion that the data from online experiments are just as interpretable as data from the laboratory. Moreover, when it comes to mTurk, studies indicate that data gathered using this platform are at least as reliable as data collected in conventional ways and that participants are even slightly more diverse than participants in standard internet studies, making it possible to generalise the results (e.g., Buhrmester, Kwang, & Gosling, 2011). In the present study, in order to establish construct validity, participants first completed the "sparks" creativity test (the same one that had been used in Study 2, completion time: 15 minutes), a test assessing abstract logical reasoning (for testing discriminant validity; completion time: 6 minutes), and a questionnaire for biographical data that included self-assessed drawing ability (in order to establish whether test performance was really independent of the ability to draw). Completing the entire battery took about 30 minutes, and participants were compensated with US$ 5 each for their participation. In order to establish test-retest reliability of the creativity test, participants completed this instrument for a second time two weeks later.

Like in Studies 1 and 2, there was no ethics commission involved for giving ethical approval for the study. However, participants were told that their participation in this study was voluntary, that it would be anonymous, and that they could withdraw from the study at any time.

Measures

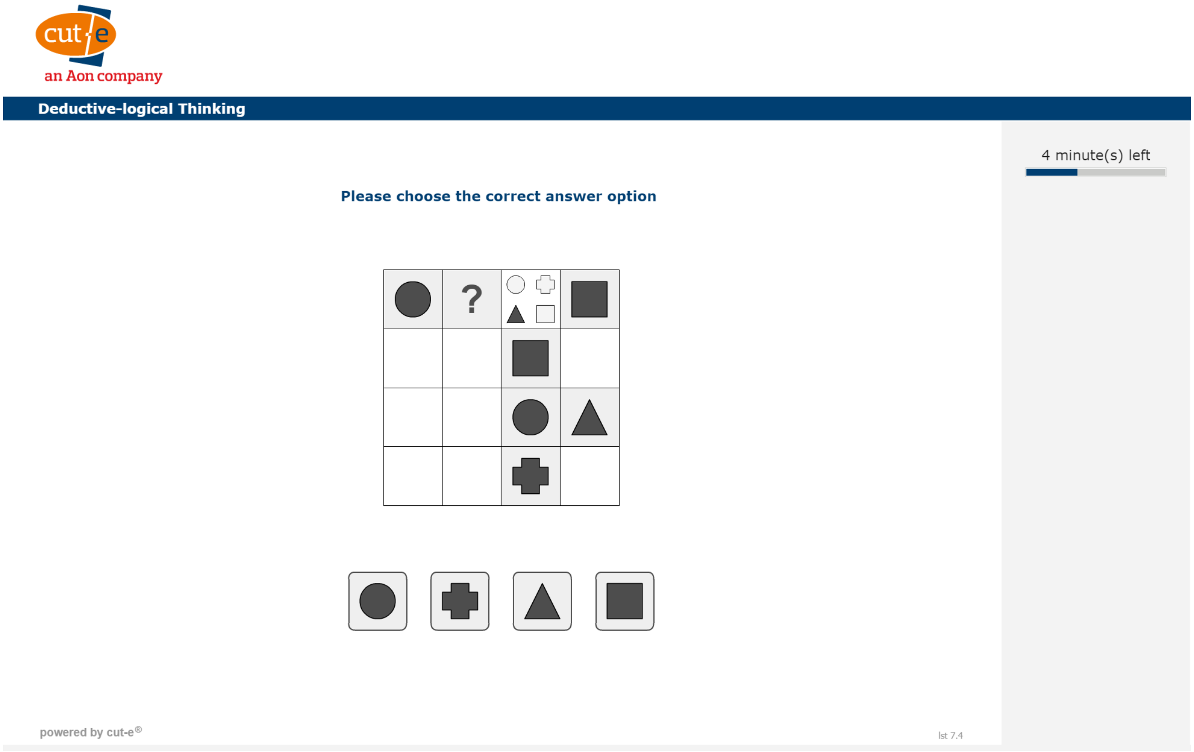
Apart from the already described creativity test, we used a measure assessing deductive-logical reasoning, scales lst (cut-e, 2008), a questionnaire assessing demographical data and a self-rating of the ability to draw. The creativity test is described in the Measures Section of Study 2. It was scored using the automated scoring algorithm and an Expert rating. The following sections thus describe the questionnaire assessing biographical data, the test assessing logical reasoning, and the rules the Expert rating of the creativity test was based on.

Biographical data. In the biographical questionnaire, candidates were asked to indicate their age, gender, and education level. Additionally, they were asked to rate their ability to draw on a five-point Likert scale (1=very poor; 5=very good).

Deductive-logical reasoning. The measure assessing deductive-logical reasoning, scales lst (cut-e, 2008) was used to establish discriminant validity. According to Guilford's (1950) theoretical assumptions on creativity as divergent thinking, we would expect moderate correlations between divergent and convergent thinking.

Scales lst (cut-e, 2008) consists of grids of 4x4 or 5x5 cubicles that contain different symbols, each of which must appear only once in each row and each column, like in Sudoku. On the test, incomplete grids are depicted in which one cubicle contains a question mark. The participant has to find the symbol to replace the question mark and select it by clicking on one of the four or five alternatives provided. The test has a split-half reliability of *a* = .89 (Spearman-Brown corrected; *N* = 3,216) and a correlation of *r* = .48, *p* < .01 (*N* = 90) with Raven's Advanced Progressive Matrices (APM; Raven, Raven, & Court, 1998).

Figure 3. Screen shot of an item of scales lst.



Please choose the correct answer option

●	?	○ ✚ ▲ □	■
		■	
		●	▲
		✚	

● ✚ ▲ ■

Expert rating of the creativity test. The online platform on which all instruments ran allowed for generating a report for the creativity test that depicts, for each participant, the objects they have drawn along with the names they gave these objects. For those who had completed the creativity test twice, this report was retrieved from the system and two trained evaluators (Experts) rated the objects based on the instructions given in the TTCT manual. Since the Expert rating was to reflect the rating a trained Expert would give using the instructions in the TTCT manual, not the rules the algorithm had been programmed on, the TTCT criteria were used. The hypothesis was that the results would be similar since the algorithm had been based on assumptions regarding the properties of picture and title drawn from the TTCT manual. Moreover, the Experts rated Originality and in addition to Study 2 also Fluency and Flexibility. Thus, the rating criteria the Experts used when rating creativity were as follows:

- **Fluency:** Number of real-world objects drawn and named (like in Study 1).
- **Flexibility:** each object was assigned to one of the categories given in the TTCT manual; for each category represented in their objects participants received one point; if it was not possible to assign an object to a category the participant received an additional point for each non-assignable object (minimum score was 1, with no theoretical maximum here)
- **Originality:** this was a combination (unweighted addition) of two TTCT categories, elaboration, i.e., complexity of the picture, and abstractness of the title, i.e., complexity of title (the possible range for Originality was from 0 up to 30 per picture).
  - elaboration: participants received one point for each non-necessary detail per object (e.g., window in house was one point, chimney on roof of house one more); scores ranged from 0 to 20 per object
  - abstractness of title: participants received 1 point for each word they added to the title of the object; if it was only one word the score was 0; each additional word (except for prepositions and articles) scored 1 point (e.g., dog was 0 points, dangerous dog 1 point, man running away from dangerous dog 3 points); scores ranged from 0 to 10 points

The two Experts rated the pictures independently of each other, their ratings were then combined by averaging them (see below in the Results section).

Results

In this section, we will first focus on the validation of the scoring algorithm by comparing its results to the ratings two trained Experts gave, using the small sample of N = 120. Next, results on test-retest reliability will be reported, also using the same sample of N = 120, and finally the results on construct validity will be presented, using the large sample of N = 470. Table 3 depicts descriptive statistics for the two samples as calculated by the scoring algorithm. Table 4 depicts the ratings on the same three dimensions as given by the Trained Experts.

Table 3
Descriptive Statistics for Fluency, Flexibility, and Originality for the Full Sample, and for the Small Sample at First and Second Test Completion as scored by the Algorithm

	Full sample		**Small sample T1**		**Small sample T2**	
	M	***SD***	***M***	***SD***	***M***	***SD***
Fluency	10.07	4.53	9.22	5.40	9.66	5.25
Flexibility	7.26	2.72	6.60	2.54	7.12	2.85
Originality	181.14	85.94	169.67	88.38	183.89	62.67

Note. Full sample: N = 470; small sample T1 (at time 1, first test completion): N = 120; small sample T2 (at time 2, second test completion for test-retest reliability): N = 120.

Table 4
Descriptive Statistics for the Fluency, Flexibility, and Originality for the Small Sample as rated by the Trained Experts (Averaged)

	M	***SD***
Fluency	8.39	4.05
Flexibility	5.98	2.47
Originality	171.74	80.71

Note. N = 120; small sample at T2 (at time 2, second test completion)

Validating the scoring algorithm

In order to compare the scoring algorithm to the Expert ratings an automated reporting function had recorded the drawings and titles so that after test completion they could be downloaded from the data base. As mentioned above these pictures were scored according to the TTCT manual by two trained Experts who did not know the results of the automated scoring.

For assessing interrater reliability, i.e., agreement among the two raters, the intra-class correlation coefficient was calculated. As a model, two-way random was chosen since the same raters rated all participants, and the type used was absolute agreement. The ICC was .80, *p* <.01 for Fluency; .93, *p* < .01 for Flexibility; and .81, *p* < .01 for Originality. According to Cicchetti (1994), this can be interpreted as excellent. Therefore, the two ratings were averaged to have an overall rating of the three scores that could be correlated with the automated score. Table 5 shows the correlations between the automated sparks scoring and the manual scoring by the two Experts.

Table 5
Correlations Between the Three Creativity Scores as Calculated by the Algorithm and as Rated by the Trained Experts Based on the TTCT Manual Using the Data from T2.

	r
Fluency	.99**
Flexibility	.85**
Originality	.88**

Note. *N* = 120. ** *p* < .01.

The results demonstrate that with .99 the automated sparks scoring almost perfectly predicts the expert rating of Fluency. This means that the scoring algorithm detects whether someone has added a shape to the one already given and whether a valid name was given to the object and thus reflects a decision made by a trained rater. The other correlations indicate that the automated scoring reliably reproduces an expert scoring according to the TTCT manual.

Assessing test-retest reliability

Test-retest reliabilities for the automated scoring (*N* = 120) were .82 for Fluency, .67 for Flexibility, .and 71 for Originality. This can be considered satisfactory.

Construct validation

For establishing discriminant validity, the creativity test scores as calculated by the scoring algorithm were correlated with the test scores of scales lst, the test assessing deductive-logical reasoning. Table 6 depicts the results.

Table 6
Intercorrelations Between the Four Creativity Scores and a Test Assessing Logical Reasoning

	Reasoning test speed	Reasoning test accuracy	Reasoning test performance
Fluency	.33**	- .17**	-.01
Flexibility	.85**	.02	.15**
Originality	- .03	.11*	.15**

Note. *N* = 470. * *p* < .05. ** *p* < .01. Reasoning test speed: Number of items processed. Reasoning test accuracy: Percentage of correct solutions of items processed. Reasoning test performance: Overall performance score on logical reasoning test.

There is no correlation between the Fluency score and the performance score of the reasoning test. The other two creativity scores and the performance score of the reasoning test is *r* = .15. This can be interpreted as an indicator for the divergent validity of the creativity test. All correlations are highly significant (*p* < .01), but according to Cohen’s (1992) taxonomy this is a small effect, with only about 2% in the variance in Flexibility or Originality, being explained by the reasoning test.

There is a rather high correlation of *r* = .33 between Fluency in the creativity test and processing speed (number of items processed) in the reasoning test, which is in line with expectations since Fluency (number of pictures generated) may also be interpreted as an indicator for a participant’s processing style. Thus, someone who works quickly on one instrument is also likely to work quickly on the other one. There is also a significant negative correlation (*r* = –.17) between Fluency in the creativity test and accuracy (percentage of correct solutions) in the reasoning test. The effect is small but the direction is to be expected: someone who works quickly is likely to not be quite as precise as someone who works more slowly. Overall the results suggest discriminant validity of the creativity test.

Finally, candidates' results as scored by the algorithm correlated almost 0 with their ability to draw (r = -.05 for Fluency, r = .07 for Flexibility, r = -.01 for Originality). Thus, performance on the test can be considered to be independent of the ability to draw.

Discussion

The study investigated the validity of the scoring algorithm and the psychometric properties of the test that was developed based on the previous two studies. The scores obtained by the automated scoring algorithm correlate highly with Expert ratings that are based on the TTCT manual instructions (between r = .85 and r = .99), thus the scoring algorithm can be considered a valid one. Test-retest reliabilities are between .67 and .82 for the three creativity scores and can be considered satisfactory. Intercorrelations with a test assessing logical reasoning are as to be expected: low but positive. Thus, discriminant validity can be shown as well. Finally, results are independent of the ability to draw.

Discussion

This paper describes how, in three studies, an online creativity test was developed that is suitable for unsupervised online assessment and that uses an automated scoring algorithm, so that it is not required to have trained experts to evaluate participants' test results.

The test is based on the well-established and validated concept of Torrance Tests of Creative Thinking (Torrance, 1974). It was adapted to the computer screen using the same logic. In Study 1, parameters that can be used for designing an automated scoring algorithm were explored. The parameters that were found to be promising were for Fluency: number of real-world objects drawn and names; for Flexibility: number

of different patterns used when combining the shapes; for Originality: uniqueness of the picture (uniqueness of the combinations of shapes), complexity of the picture (number and diversity of shapes used, degree of variation of the shapes in terms of size, position and rotation), uniqueness of the title, and complexity of the title (number of words used and length of each word). In Study 2, this scoring algorithm was validated by comparing the results from this algorithm to Expert ratings. Finally, in Study 3, the final version of the test was validated and showed satisfactory overlap with Expert ratings, test-retest reliability, discriminant validity, and independence of the ability to draw.

Limitations

A limitation of Study 1 and thus, in the definition of the scoring algorithm, is that the development of the algorithm was not theory-driven, but rather based on what was available. When the algorithm was developed, the authors could not find any models or studies on the aspects of the objects created that actually turn them into more or less creative ones. Thus, the method used to develop the algorithm was a machine-learning one to transfer the rating knowledge of experts into an algorithm. However, the subsequent two studies showed that the scoring is valid. Nevertheless, further validation of the scoring is necessary. The first question will be to what extent the algorithm generalises across different samples with different educational or cultural backgrounds. Moreover, there could also be differences in the validity of the algorithm that originate from different levels of experience with using the computer or even from different personality characteristics. For example, candidates who are more experienced in using the computer and who work with graphics a fair bit might use more of the functionalities the test offers than less experienced individuals. If true, the scoring algorithm might underestimate the creativity of those individuals who do not play around with the shapes quite as much.

Another issue of the current algorithm is that it does not check whether the object and its name are really in line. What could ultimately happen is that participants simply pull random shapes onto the drawing area and assign names that have nothing to do with the objects. To account for this, a report option was created in which the assessor can see all the objects and the names assigned to them. Thus, evaluators can manually check for congruence between object and name in case of doubt. Moreover, one could also state that, even if there is no congruence between objects and given names, under certain circumstances the individual acting in such a way can still be considered creative: In order to get a high score on originality they will still have to come up with various combinations of objects and they will have to come up with many different names. So even if there is no congruence between object and name they will have had a lot of associations and a wide range of different ones.

In general, what is still lacking is proof of the prognostic validity of the instrument. There needs to be a criterion-related validity study in which it is tested whether the instrument predicts creativity. This is currently being planned.

Implications

Sparks assesses creativity and can be used in unsupervised settings. Thus, it can be used to assess an individual's creativity either at the selection or at the development stage. Creativity, in turn, is an important building block of innovation. Hence, if an organisation is looking into recruiting or developing individuals who are likely to be workplace innovators it can use this instrument.

At the selection stage, the recruiting organisation can determine who is creative and therefore might come up with ideas for workplace innovation when being recruited. Since the test is optimised for unsupervised online assessment this can happen at a very early stage in the recruitment process, i.e., before the candidate is even invited to an on-site interview or assessment centre. This means that it can be done quickly, at low cost, and independently of where on the globe the candidate is located.

At the developmental stage, the instrument can be used for assembling and developing teams as well as individuals. In combination with instruments assessing logical reasoning and personality, organizations can assess who in the team is likely to detect structures or processes within the organisation that can be optimised (based on a test assessing analytical abilities), who will come up with creative ideas for this issue (based on sparks), and who will be able to implement and communicate the innovation well (based on personality characteristics such as conscientiousness or social confidence). Thus, teams can be assembled in the best possible way so that they can be innovators in the workplace. Thus, altogether, sparks can help select and develop potential workplace innovators.

References

Amabile, T. M., Conti, R., Coon, H., Lazenby, J., & Herron, M. (1996). Assessing the work environment for creativity. *Academy of Management Journal*, 39(5), 1154-1184.

Bateson, J., Wirtz, J., Burke, E., & Vaughan, C. (2013). When Hiring, First Test, and Then Interview. *Harvard Business Review*, 91(11), 34.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon‘s Mechanical Turk. A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6(1), 3-5.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4), 284-290. doi:10.1037/1040-3590.6.4.284.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

cut-e (2002). *shapes.* Hamburg: cut-e Group.

cut-e (2008). *scales lst.* Hamburg: cut-e Group.

cut-e (2012). *The cut-e Assessment Barometer 2012/13: The Global Survey of Psychometric Assessment Usage.* Hamburg: cut-e Group.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological Monographs*, 58(5), 1–113.

Farr, J. L., Sin, H., & Tesluk, P. E. (2003). Knowledge Management Processes and Work Group Innovation. In L. V. Shavinina (Ed.), *International Handbook on Innovation* (pp. 574-586). Ottawa: Elsevier Science Ltd.

Guilford, J.P. (1950). Creativity. *American Psychologist*, 5, 444-454.

Guilford, J.P. (1967). *The Nature of Human Intelligence.* New York: McGraw Hill.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. Form 4. In W. Sarges & H. Wottawa (Eds.), Handbuch wirtschaftspsychologischer Testverfahren* (pp. 95-101). Lengerich: Pabst Science Publishers.

Kesselring, A., Blasy, C., & Scopetta, A. (August, 2014). *Workplace Innovation: Concepts and indicators. Brussels: European Commission*, DG Enterprise and Industry.

Kim, K. H. (2006). Can We Trust Creativity Tests? A Review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, 18(1), 3-14.

Lochner, K. (2016). *Successful Emotions. How Emotions Drive Cognitive Performance.* Wiesbaden: Springer.

Maier, G. W., Streicher, B., Jonas, E., & Frey, D. (2007). Innovation und Kreativität. In D. Frey & L. von Rosenstiel (Eds.), *Enzyklopädie der Psychologie: Wirtschaftspsychologie* (pp. 809-855). Stuttgart: Hogrefe.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The Integrity of Web-Delivered Experiments: Can You Trust the Data? *Psychological Science*, 11(6), 502–506. doi:10.1111/1467-9280.00296

Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R.* Göttingen: Hogrefe.

Raven, J. C., Raven, J. & Court, J. H. (1998). *Matrizen-Test-Manual Band 2.* Göttingen: Beltz-Test GmbH.

Torrance, E. P. (1966). *The Torrance Tests of Creative Thinking – Norms-Technical Manual Research Edition. Princeton,* NJ: Personnel Press.

Torrance, E. P. (1974). *The 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.

Torrance, E. P. (1981). Empirical validation of criterion-referenced indicators of creative ability through a longitudinal study. *Creative Child and Adult Quarterly*, 6, 136-140.