The Alternative Uses Task: A Comparison between various Scoring Methods

Charlotte Tanis

University of Amsterdam

Word count: 4834
Student number: 10304533
Bachelor Project
Claire Stevenson
03-07-2017

Abstract

The Alternative Uses Task (AUT) is a frequently used tool to assess creativity. To score the AUT, a new algorithm-based consensual assessment technique was developed. This algorithm and two traditional methods (uniqueness and expert judges) were evaluated in terms of reliability and validity. Data consisted of an available set (*n* = 303) and a new set was obtained (*n* = 157). Uniqueness scoring showed a robust relationship with fluency, strong internal consistency, and moderate to strong test-retest reliability. Judges scores had a weak relation with fluency, a moderate internal consistency, strong inter-rater reliability, and weak to moderate test-retest reliability. The algorithm was not associated with fluency and had a moderate internal consistency. All methods showed either no or weak relation with CAQ and IWB. None of the methods showed consistently adequate qualities for reliable and valid testing. However, the algorithm shows potential and further improvement is attainable through data cleaning and expansion of the reference database.

Keywords: creativity, divergent thinking, Alternative Uses Task, scoring methods

The Alternative Uses Task: A Comparison between various Scoring Methods

Creative ideas come in a wide variety of shapes and forms, making it difficult to judge how creative they truly are. According to the most commonly used definition, there are two broad criteria a creative idea should satisfy (Runco & Jaeger, 2012). First, for something to be creative it has to be original. This component may be unsurprising since a common idea can hardly qualify as creative after all. Though crucial, originality alone is insufficient; a creative idea should be useful as well. Stein (1953) was the first to define creativity as a combination of originality and utility. According to Stein, “the creative work is a novel work that is accepted as tenable or useful or satisfying by a group in some point in time” (p. 311). This study is a comparison of various scoring methods of the Alternative Uses Task and an evaluation of their suitability in measuring creativity.

To adequately examine creativity, assessments should have high reliability and validity, and, preferably, be user-friendly (Baas & Van der Maas, 2015). The reliability of a test indicates to what extent the measurements represent true values. Reliability can be interpreted as the variance in scores introduced by differences between participants as opposed to variance introduced by error. When the reliability of a test is low, it is unclear where differences in scores stem from and test scores become meaningless. Reliability of a test is estimated by the test-retest reliability, inter-rater reliability (IRR), and internal consistency (Cook & Beckman, 2006). Test-retest reliability shows the extent to which test scores are predictive of subsequent scores from the same test. The internal consistency is a measure of the relationship between different items of a test. A high internal consistency indicates that various items of a test measure the same concept. IRR indicates the degree to which different judges agree on the score and is therefore only relevant when judges do the scoring.

High reliability indicates low measurement error, however, it does not guarantee the testing of the intended measure. This is where validity comes into play. It is impossible to determine with a direct evaluation if a test is valid, but evidence can be gathered to indicate that the interpretations of the results are supported (Downing, 2003). For the divergent validity of a test, there should be no correlation with tests that do not measure creativity, for example, a social desirability questionnaire. A high predictive validity is achieved when a test of creativity also shows accuracy in predicting creative behavior. Additionally, scores on a creativity test should be correlated with other measures of creativity to ensure an appropriate convergent validity. It is also possible to look at other constructs that have been shown to be related to creativity. For instance, a large body of research has found a correlation between creativity and openness to experience (Silvia et al., 2008).

There are many measures of creative potential, behavior, and achievement (Baas & Van der Maas, 2015). Creative potential is frequently measured with divergent thinking tasks (Runco, Abdulla, Paek, Al-Jasim, & Alsuwaidi, 2016). A widely used divergent thinking task is the Alternative Uses Task (AUT) designed by Guilford (1967). In this task, participants are instructed to come up with as many creative uses as possible for a common object. A brick, for example, can be used to build a house, but also to secure one's tent while camping, or could be worn as a highly uncomfortable hat. Different methods are available to score the AUT, each with their own merits.

The most simple method bases the score on the fluency of ideas, measured by the number of responses per participant. This objective measure is both fast and applied with little difficulty, but fails to take the originality or usefulness of ideas into account. A participant who lists many common uses will receive a higher score than one who lists a few highly creative responses. This illustrates that the quality, instead of the number, of responses reflects the creative potential of a participant. It is, therefore, argued that originality and fluency should be two distinct factors. To ensure the discriminant validity of the AUT, its scores should not be related to the fluency measure (Silvia et al., 2008). Differences in scores could otherwise (for a large part) be explained by variation in fluency, instead of true differences in creativity. However, the results of many commonly used scoring methods are currently influenced by the fluency of ideas (Plucker, Qian, & Wang, 2011).

A different objective scoring method rates originality by the statistical uniqueness of responses. One way to score uniqueness is to start utilizing a points system; for example a point can be given for every response that only occurs once in the sample and no points for all other responses. For every participant, the points of their responses are counted to obtain the final score. A drawback of this method is that the sample size has a direct influence on the final scores (Silvia et al., 2008). Another way of scoring uniqueness assigns a point to responses given by less than five percent of the sample (Milgram & Milgram, 1976). This diminishes the effect of sample size on scores, but other complications remain. For example, an uncommon response is not necessarily a creative response. Both responses that are extremely creative but without practical use, and responses containing such common uses that only very few people will take the time to write down will be rewarded (Silvia et al., 2008). Furthermore, there is a strong relation between fluency scores and uniqueness scores, so an adequate discriminant validity is not accomplished (Silvia et al., 2008).

To overcome these issues, the consensual assessment technique (CAT), can be used instead. This technique, first described by Amabile (1982), utilizes expert evaluations to score the creativity of responses. When applied to the AUT, each response receives a creativity score from several expert judges, and these scores are averaged to get the creativity score of a single response (Silvia, 2011). Most commonly, for each participant the scores of all answers are averaged to calculate the final score. This score is not dependent on the number of ideas, but merely looks at the quality of ideas submitted. To ensure the reliability of this method a high inter-rater consistency is required between judges. Kaufman, Lee, Baer and Lee (2007) showed that at least five raters and 15 items are needed to expect an agreement between judges above .80 when judging the creativity of photograph captions. A drawback of this time intensive method is that the final score is averaged from all given responses. Hence, participants with two highly creative responses will receive a higher score than someone who has given two responses of the same quality, but also submitted two moderately creative responses. One could easily argue that the latter should receive a higher score.

A novel method to score divergent thinking tasks depends on an algorithm. Beketayev and Runco (2016) used a semantics-based algorithm to score the Many Uses task, which only differs from the AUT in the objects used. Fluency, flexibility and originality scores were obtained utilizing a more traditional scoring method as well as an algorithm based score. Traditional flexibility scores were calculated by counting the categories of responses used, and originality scores were based on the uniqueness of answers. The algorithm used semantic networks to calculate the number of categories for the flexibility score. Originality scores were based on the distance between a participant's responses in the semantic networks. There was a strong relationship between the flexibility scores produced by the traditional method and the algorithm. The originality score of the algorithm, however, was only weakly related to the traditional originality score. Originality scores of the algorithm were based on existing norms, instead of comparing responses. The authors concluded that the semantic based algorithm works well to investigate ideational flexibility, but had some drawbacks when measuring the originality of ideas.

An automatic scoring method such as the algorithm used by Beketayev and Runco has great benefits. No judges are required to score the test, making the scoring process more cost and time efficient. This reduction in cost and time may persuade more researchers to include creative potential in new studies (Silvia, Martin, & Nusbaum, 2009).

A novel large-scale consensual assessment technique in the form of an algorithm was developed for this study with the aim to accurately measure the originality of AUT responses. Scoring was based on a large database of previously obtained responses. After cleaning the data, all responses in the database were rated by expert judges on both originality and utility. New responses were matched to ones already present in the database and given a score accordingly.

By introducing this novel scoring method and comparing two others in terms of reliability and validity, this study will contribute to earlier research into the psychometric qualities of the AUT. The methods compared are uniqueness, the average of expert ratings, and the average of the algorithm's rating. Besides two to four versions of the AUT, participants will complete two Verbal Fluency tasks, the Raven progressive matrices task, the Innovative Work Behavior scale, and part two of the Creative Achievement Questionnaire.

By not scoring the usability of responses this study can not definitively address creativity as a whole, however, it could be a valuable addition to current discourse. The primary goal is to assess if an algorithm-based scoring method can yield equally reliable, if not more reliable, results as scoring by judges. If this is proven to be the case, not only would this call for more in-depth research into algorithm based scoring methods, it provides potential for low-cost creativity testing.

Methods

Participants

The design of this study is based on the use of two data sets. The first set was collected in 2016 and will hereafter be referred to as "data 2016". The second dataset was collected during the bachelor project for which this thesis was written, and is specified as "data 2017". The only inclusion criterium for both sets was an age between 17 and 25. A total of 10 and 14 participants were excluded from data 2016 and 2017 respectively for not matching the criteria.

After corrections, participants of data 2016 consisted of 303 first year Psychology students (91 male, 212 female, M_{age} = 20.1 years, age range: 17-25 years) from the University of Amsterdam. Students were required to participate to complete their first year.

Participants of data 2017 consisted of 157 Dutch students (88 male, 69 female, M_{age} = 19.8 years, age range: 17-24 years). Students were recruited from the Nova College in Hoofddorp (trade school, *n* = 97), the Faculty of Business and Economics of the Amsterdam University of Applied Sciences (*n* = 30), and the Faculty of Science of the UvA (*n* = 30). One humanities, 125 social science, and 31 science students took part in this study. All participants received a compensation of €10. This research was approved by the local Ethics Review Board prior to data collection.

Materials

During the collection of data 2017, participants completed part two of the Creative Achievement Questionnaire (CAQ), two to four AUT's, a 42 item Raven Progressive Matrices task (RPM), two Verbal Fluency tasks (VF), and the Innovative Work Behavior scale (IWB).

Part two of the CAQ was used to measure creative achievement. In this test, concrete achievements were listed in ten domains such as visual arts and scientific discovery. Participants were asked to indicate which of the achievements apply to them. Each domain contained eight achievements for which zero to seven points are rewarded. It was allowed to check more than one achievement per domain. To illustrate, items in the domain “visual arts” worth 0, 1, 4, and 7 points respectively were: “I have no training or recognized talent in this area. (Skip to Music)”, “I have taken lessons in this area”, “I have had a showing of my work in a gallery”, and “My work has been critiqued in national publications”. To calculate the total creative achievement score all points were added up, leading to a minimum score of 0 and a maximum score of 280. The CAQ has a high test-retest reliability (r = .81), high internal consistency (α = .96), and is correlated with other measures of creativity (Carson, Peterson, & Higgins, 2005).

The items used in the AUT (Guilford, 1967) were a brick, fork, paperclip and towel. Participants were instructed to name as many possible creative uses for these objects within a timeframe of two minutes per object. Paperclip and towel were added after part of data 2017 was already collected, as participants completed the tasks quicker than initially presumed.

The RPM (Raven, 2000) consisted of 24 even numbered Standard Progressive Matrices items, and 18 even numbered Advanced Progressive Matrices items. The items used in the VF tasks (Hills, Jones, & Todd, 2012) were animals and occupations. For both VF tasks, there was a time limit of one minute in which participants had to write down as many types of animals and jobs as possible. Both the RPM and VFs were used to collect data for different studies, and will therefore not be further discussed in this article.

The IWB scale is a test developed by Janssen (2000) to measure innovative work behavior. Innovative work can be defined by three aspects, namely idea generation, idea promotion, and idea realization. The IWB scale contained nine items in total, and three items per aspect, i.e. “I often create new ideas for difficult issues” (idea generation). Respondents indicated how often they perform each behavior on a seven-point Likert scale (1 = never, 7 = always). For each item, a score of one to seven points was given, corresponding to the numbers on the Likert scale. The average score of the nine items formed the total IWB score. Janssen (2000) reports correlations between the different aspects of innovative work behavior ranging between .76 and .85, and a high internal consistency (α = .96).

During the collection of data 2016, participants also completed the CAQ, four AUT’s and IWB. Since this was a large scale data collection, other tasks for different studies were also included, which will not be discussed in this article.

Procedure

The collection of data 2017 took place in classrooms at schools and universities where participants were recruited. Different sessions were held where groups of students took part in the study simultaneously. Students completed all tasks online using their laptop. The time limit for the experiment was one hour, after which the experiment automatically stopped.

After giving informed consent, a few of basic questions such as age, gender, and field of studies followed. Participants then completed the tasks in the following order: AUT (brick, fork, paperclip, towel), VF (first animals, then jobs), RPM, and CAQ. After these tasks, participants were shown AUT responses from data 2016 and judged whether these were creative or not. Lastly, participants indicated how important (“not at all important”, “not important”, “neutral”, “important”, or “very important”) usability, innovation, suitability, and originality each were in judging creativity. After completion of the tasks, participants received the compensation in cash.

Data 2016 was collected in a large computer classroom at the UvA. Participants took part in multiple sessions of data collection spread over several days. Among a large number of other tasks were the AUT's, CAQ, and IWB. The four objects of the AUT were divided over different sessions. Brick and fork were part of an earlier session, and a subset of students also completed paperclip and towel during a later session. All tasks were completed on computers available in the classroom.

Originality scoring of the AUT

Scoring according to the uniqueness method rewarded one point for every response given by less that five percent of the sample. All other responses received no points. For the final score all points were added.

To generate the average of the expert ratings each response was scored for originality by two expert judges. Possible scores were one (not at all original) to five (very original). If a response was nonsensical, the response was scored zero. For every response, the average score of the two judges was calculated if both judges gave a non-zero score. If one judge gave a score of zero, while they other did not, the non-zero score was used. If both judges gave a score of zero, the response was excluded from the final score. The final score for each participant was calculated by taking the average of the scores of the participant's responses.

A database of unique responses per object was created to produce a score by the algorithm. All responses from data 2016 were cleaned, i.e. spell-checked, punctuation and stopwords removed, and stripped of white space. The database for every object contained all unique cleaned responses. Every unique response was given a score per expert judge by taking the median of the judge's scores of that response. To calculate the score per response, the average was taken of the median judge scores if both judges had a non-zero median. This average was rounded to the nearest integer. If one judge had a non-zero median and the other did not, the non-zero median was used. If the median of both judges was zero, the response was deleted from the database.

All responses from data 2017 were cleaned following the same steps taken while creating the reference database. The cleaned responses were then matched with the database. If the response could be located, it received the score from the database. If no match could be made, the response did not contribute to the total score. The average of matched responses formed the final score.

Statistical Analysis

Several analyses were applied for each measurement of originality discussed above. These analyses were conducted on data 2016 as well as data 2017, with the exception of the algorithm. Due to the algorithm utilizing data 2016 as its database, it was impossible to analyze this data set using the algorithm. Therefore, the analyses for the algorithm were only conducted on data 2017.

First, the correlation between fluency and originality scores on a participant level was computed. Second, a reliability analysis was conducted in terms of internal consistency, IRR, and test-retest reliability (where possible). Internal consistency was determined by the correlation between originality scores on participant level of two different objects utilizing the same scoring method. This was calculated when data of different objects was collected during the same session. The intra class coefficient was used to calculate the IRR for expert ratings on response level. The test-retest reliability was determined by the correlation between originality scores on participant level of objects when the data was collected during different sessions. Lastly, convergent validity was assessed by the correlation between originality scores and CAQ scores, and between originality scores and IWB scores, both on participant level.

All correlations were calculated with the cor.test function from the R *stats* package (R Core Team, 2016). The intra class coefficient was calculated with the icc function from the R *irr* package (Gamer, Lemon, & Fellows, 2012).

Results

In both data 2016 and 2017 not every task was completed by all participants. Table 1 shows the number of participants per object of the AUT. For Data 2017 the table also displays per object how many participants completed both AUT and IWB. Unfortunately, the data of the CAQ was corrupted and could not be used for analysis. Participants in data 2016 all completed both IWB and CAQ. All available data was used for analysis.

Table 1

Number of participants who completed an object or object combination of the AUT in data 2016 (n = 303) and data 2017 (n = 157).

		Data 2017: *n*	
	Data 2016: *n*	AUT	AUT + IWB
AUT			
Brick	285	153	148
Fork	295	155	151
Paperclip	192		
Towel	98		
Brick and Fork	185	151	
Paperclip and Towel	96		
All objects	90		

Table 2 (data 2017, brick and fork), 3 (data 2016, brick and fork), and 4 (data 2017, paperclip and towel) show the correlations between originality scores produced by different scoring methods of the AUT. These tables show the correlations within the object, and between objects for data collected during the same session.

Table 2

Data 2017: Brick and Fork. Intercorrelations for fluency, uniqueness, judges, and algorithm scoring.

			Brick				Fork			
Measure	*M*	*SD*	1	2	3	4	5	6	7	8
Brick										
1. Fluency	6.26	3.21	—							
2. Uniqueness	4.20	2.50	**.88 [.83, .91]**	—						
3. Judges	1.95	.64	-.06 [-.21, .10]	**.21 [.05, .36]**	—					
4. Algorithm	1.60	.48	.01 [-.15, .17]	**.20 [.04, .36]**	**.65 [.55, .74]**	—				
Fork										
5. Fluency	6.32	3.32	**.71 [.62, .78]**	**.54 [.41, .64]**	-.12 [-.28, .03]	-.08 [-.24, .08]	—			
6. Uniqueness	4.30	2.72	**.65 [.55, .74]**	**.62 [.51, .71]**	.09 [-.07, .24]	.09 [-.08, .25	**.88 [.84, .91]**	—		
7. Judges	2.23	.69	-.15 [-.30, .01]	.13 [-.02, .29]	**.48 [.35, .60]**	**.39 [.24, .52]**	**-.24 [-.38, -.08]**	.03 [-.12, .19]	—	
8. Algorithm	1.90	.73	-.13 [-.29, .03]	.06 [-.11, .22]	**.49 [.36, .61]**	**.40 [.24, .53]**	-.10 [-.26, .06]	.12 [-.04, .28]	**.69 [.59, .77]**	—

Note. Numbers in brackets are 95% confidence intervals of the correlation coefficients. All bold coefficients are significant at p < .05.

Table 3

Data 2016: Brick and Fork. Intercorrelations for fluency, uniqueness, and judges scoring.

			Brick			Fork		
Measure	*M*	*SD*	1	2	3	4	5	6
Brick								
1. Fluency	11.42	7.52	—					
2. Uniqueness	7.00	5.63	**.91 [.89, .92]**	—				
3. Judges	1.88	.46	**.12 [.01, .23]**	**.30 [.19, .40]**	—			
Fork								
4. Fluency	13.46	8.15	**.66 [.59, .72]**	**.57 [.49, .64]**	.00 [-.12, .12]	—		
5. Uniqueness	8.86	6.58	**.63 [.55, .69]**	**.61 [.53, .68]**	**.12 [.01, .23]**	**.93 [.92, .95]**	—	
6. Judges	2.21	.50	-.04 [-.16, .07]	.10 [-.02, .21]	**.57 [.48, .64]**	**-.16 [-.27, -.05]**	.01 [-.11, .12]	—

Note. Numbers in brackets are 95% confidence intervals of the correlation coefficients. All bold coefficients are significant at p < .05.

Table 4

Data 2016: Paperclip and Towel. Intercorrelations for fluency, uniqueness, and judges scoring.

			Paperclip			Towel		
Measure	*M*	*SD*	1	2	3	4	5	6
Paperclip								
1. Fluency	7.58	3.74	—					
2. Uniqueness	6.31	3.48	**.95 [.94, .87]**	—				
3. Judges	2.54	.45	**.16 [.02, .29]**	**.17 [.03, .30]**	—			
Towel								
4. Fluency	10.07	4.69	**.67 [.54, .77]**	**.64 [.51, .75]**	-.07 [-.27, .13]	—		
5. Uniqueness	6.77	4.51	**.65 [.52, .76]**	**.65 [.52, .76]**	-.10, [-.29, .10]	**.89 [.85, .93]**	—	
6. Judges	2.25	.40	-.12 [-.31, .08]	-.15 [-.34, .05]	**.53 [.37, .66]**	**-.29 [-.46, -.10]**	**-.26 [-.43, -.06]**	—

Note. Numbers in brackets are 95% confidence intervals of the correlation coefficients. All bold coefficients are significant at p < .05.

The algorithm was able to match part of the responses with the database and gave a score accordingly. Participants gave a total of 1061 responses for brick and 1089 for fork in data 2017, of which 683 and 664 were matched respectively.

The correlation between the scoring methods and the fluency measurement was computed within the object to assess the discriminant validity. Uniqueness was strongly related to fluency (*r* = .88 to *r* = .95). For the expert ratings, these correlations range from very weak to weak (*r* = -.29 to *r* = .16). The algorithm scores were only computed for data 2017, and no significant correlations were found between these scores and fluency.

To compute the internal consistency, the correlation between a participant's total scores based on the same scoring method applied to different objects used in the same session was computed. Both fluency and uniqueness show a strong relation between the scores of different objects (*r* = .66 to *r* = .71 for fluency, and *r* = .63 to *r* = .65 for uniqueness). However, both expert ratings and the algorithm scores only show a moderate correlation (*r* = . 48 to *r* = .57 for expert ratings, and *r* = .40 for the algorithm scores).

For expert ratings, the IRR was assessed by computing a two-way mixed, agreement, average-measures intra class coefficient (ICC) for every object in the AUT. The ICC shows to what extent the different judges agree in their ratings across responses. Table 5 displays the ICC for every object. Good to excellent ICCs were found for all objects (*ICC* = .76 to *ICC* = . 97) illustrating a high agreement between judges.

Table 5

Intra class coefficients for expert ratings of the AUT to assess IRR.

	Data 2016						Data 2017					
	ICC	95% CI	*F*	*df1*	*df2*	*p*	*ICC*	95% CI	*F*	*df1*	*df2*	*p*
Brick	.76	[.74, .78]	4.25	3247	1797	<.001	.85	[.83, .87]	6.89	918	859	<.001
Fork	.85	[.82, .87]	6.76	3970	355	<.001	.97	[.96, .97]	31.6	980	981	<.001
Paperclip	.81	[.78, .83]	5.34	1454	519	<.001						
Towel	.78	[.75, .81]	4.65	986	779	<.001						

Note. A two-way, mixed, agreement, average-measures intra class coefficient (ICC) was used.

The last reliability measure investigated was the test-retest reliability. This could only be computed using data 2016 since participants in data 2017 completed all tasks in one session. During collection of data 2016, both ‘brick’ and ‘fork’ were in one session, while ‘paperclip’, and ‘towel’ were in a later session. Table 6 shows the correlations between scores on each pair of objects, in different sessions, for every scoring method. Moreover, the correlation between the mean of ‘brick’ and ‘fork’ scores and the mean of ‘paperclip’ and ‘towel’ scores is displayed. Fluency scoring had a moderate to strong test-retest correlation (r = .48 to r = .69). Both uniqueness scoring and expert ratings showed a weak test-retest relation (uniqueness: r = .36 to r = .56, judges: r = .33 to r = .49). Combining the scores of brick and fork, and of paperclip and towel resulted in higher correlations between the scores (fluency: r = .72, uniqueness: r = .57, judges: r = .57). However, these correlations still fall in the 95% confidence interval when the scores of single objects are used.

Table 6

Data 2016: Correlations between scores on different objects of the AUT taken during different sessions to assess the test-retest reliability.

	r
Brick (session 1) vs. Paperclip (session 2)	
Fluency	.59 [.49, .68]
Uniqueness	.50 [.38, .60]
Judges	.33 [.20, .46]
Brick (session 1) vs. Towel (session 2)	
Fluency	.62 [.47, .73]
Uniqueness	.36 [.17, .53]
Judges	.48 [.31, .63]
Fork (session 1) vs. Paperclip (session 2)	
Fluency	.48 [.36, .58]
Uniqueness	.50 [.38, .60]
Judges	.43 [.31, .54]
Fork (session 1) vs. Towel (session 2)	
Fluency	.69 [.56, .78]
Uniqueness	.56 [.40, .69]
Judges	.49 [.32, .63]
Mean Brick and Fork (session 1) vs. Mean Paperclip and Towel (session 2)	
Fluency	.72 [.60, .80]
Uniqueness	.57 [.41, .69]
Judges	.57 [.41, .69]

Note. Numbers in brackets are 95% confidence intervals of the correlation coefficients. All coefficients are significant at p < .05.

Convergent validity was assessed by computing the correlations between AUT scores and CAQ scores (only data 2016), and between AUT scores and IWB scores. Correlations between AUT scores and CAQ scores were not significant, very weak or weak (fluency: not significant to $r = .19$, uniqueness: not significant to $r = .26$, judges: not significant to $r = .26$). Correlations between AUT scores and IWB scores were also not significant, very weak or weak (fluency: not significant to r = .17, uniqueness: not significant to r = .23, judges: not significant to r = .15, algorithm: not significant).

Table 7

Intercorrelations for scoring methods of the AUT and the CAQ and IWB per AUT object to assess the convergent validity.

	Data 2016		Data 2017
	CAQ	IWB	IWB
Brick			
Fluency	**.19 [.08, .30]**	**.17 [.06, .28]**	.05 [-.11, .21]
Uniqueness	**.26 [.15, .37]**	**.22 [.11, .33]**	.05 [-.11, .21]
Judges	**.20 [.09, .31]**	**.15 [.03, .26]**	.03 [-.14, .18]
Algorithm			.09 [-.08, .25]
Fork			
Fluency	**.13 [.02, .24]**	**.17 [.05, .27]**	.12 [-.04, .27]
Uniqueness	**.18 [.06, .29]**	**.23 [.12, .33]**	**.18 [.02, .33]**
Judges	.11 [-.01, .22]	**.15 [.03, .26]**	.11 [-.05, .26]
Algorithm			.13 [-.03, .29]
Paperclip			
Fluency	**.19 [.05, .32]**	**.17 [.03, .30]**	
Uniqueness	**.17 [.03, .31]**	**.19 [.05, .32]**	
Judges	.07 [-.07, .21]	.03 [-.11, .17]	
Towel			
Fluency	.01 [-.18, .21]	.12 [-.08, .31]	
Uniqueness	.07 [-.13, .26]	.16 [-.04, .35]	
Judges	**.26 [.07, .43]**	.02 [-.17, .22]	

Note. Numbers in brackets are 95% confidence intervals of the correlation coefficients. All bold coefficients are significant at p < .05.

Discussion

Three originality scoring methods of the AUT, namely uniqueness, expert, and algorithm scoring, were evaluated in this study in terms of reliability and validity. For a scoring method to perform well, it should have no correlation with fluency scoring, a high internal consistency, a high test-retest reliability, and high inter-rater reliability (if applicable). Moreover, it should be related to other measures of creativity such as the CAQ and IWB.

Uniqueness scoring had a strong relation to fluency scoring, indicating that it failed to measure an aspect independent of fluency. This relation is problematic since fluency and originality are thought to be two independent factors (Silvia et al., 2008). Both the expert ratings and the algorithm scores did not have any or only a weak relation with fluency and hereby fulfilled this first criterium.

Uniqueness scoring, however, did outperform both expert ratings and algorithm scores with respect to the internal consistency. Scores of different objects were only moderately related for the latter two, meaning that the object chosen to be used in the AUT had a great impact on the scores. This is undesirable, since the creative potential of the participant should be the only influence on the score.

As shown by the high IRRs, different judges providing the expert rating agreed to a large extent. It can thus be concluded, as little variance in scores was introduced by differences between judges, that they scored consistently. While the high IRR adds to the reliability of the method, it does not provide an explanation for the insufficient internal consistency.

Both uniqueness scoring and expert ratings were shown to have a weak test-retest reliability. However, this is to be expected given their internal consistencies. When there is only a moderate relation between scores on different objects administered during the same session, it would be unexpected if this would improve when objects are administered during different sessions. The test-retest reliability did improve slightly when scores of different objects were combined. This improvement suggests that combing more than two objects may further improve the test-retest reliability. A simulation study should be conducted to test if this hypothesis holds and if so, what number of objects should be combined to expect a satisfactory test-retest reliability.

An adequate convergent validity could, unfortunately, not be demonstrated in this study. None of the scoring methods consistently showed a strong relationship with either the CAQ or the IWB. The proportion of significant relations was slightly higher for the CAQ than the IWB. The correlations between CAQ and originality scores found in this study are lower than the correlation of .49 reported in earlier research (Carson et al., 2005). In the research of Carson et al., however, the originality score was constructed by combining multiple divergent thinking tests, while in this study a single AUT was used.

The uniqueness and algorithm scoring procedures can be further improved, potentially resulting in a higher reliability for both measures. Cleaning the responses was done in the same manner for both measures. The cleaning process can be more thorough by abbreviating every word in the responses to the stem of the word, changing synonyms of words to the same default answer, and ignoring the order of words within a response.

For uniqueness scoring, further cleaning could lead to fewer responses being scored as unique since more variations of similar answers will be counted as identical, and thereby reducing the relation with fluency scores. For the algorithm scores, this could mean that more responses can be matched to the database and counted into the final score. A greater proportion of matched responses can furthermore be achieved by adding more data to the database. It is expected that increasing the matching rate will lead to an improvement in reliability.

Another suggestion for further research is the inclusion of a personality test to measure openness to experience. Due to time constrictions, it was not possible to include this measure in the current study. Earlier research has shown an association between creative potential and openness to experience (Silvia et al., 2008). If this relationship is indeed present, it could add to the validity of the methods.

A final suggestion to improve both expert ratings and algorithm scoring is to add utility to the score. Currently, the scores are solely based on originality which is essential to creativity but falls short of the complete definition, since usability is required for an idea to be truly creative (Runco & Jaeger, 2012). A possible method would rate responses on both aspects and use a weighted sum to come to a score. The exact weights which should be used could be investigated in a future study.

In conclusion, none of the methods discussed in this study fulfill all requirements an assessment should. There is, however, a clear indication that algorithm based research into creativity has great potential, as will undoubtably be further investigated in future studies.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997-1013. doi: 10.1037/0022-3514.43.5.997

Baas, M., & Van der Maas, H. L. J. (2015). De (on)mogelijkheid van een valide meting van creatief potentieel voor selectiedoeleinden. *Gedrag & Organisatie, 28*(2), 78-97. doi: 10.5553/GenO/092150772015028002002

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europes Journal of Psychology, 12*(2), 210-220. doi: 10.5964/ejop.v12i2.1127

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal, 17*(1), 37-50. doi: 10.1207/s15326934crj1701_4

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. The American Journal of Medicine, 119(2), 166.e167-166.e116. doi: http://dx.doi.org/10.1016/j.amjmed. 2005.10.036

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. Medical Education, 37(9), 830-837. doi: 10.1046/j.1365-2923.2003.01594.x

Gamer, M., Lemon, J., & Fellows, I. (2012). irr: Various coefficients of interrater reliability and agreement (Version 0.84). Available from https://CRAN.R-project.org/ package=irr

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431-440. doi: 10.1037/a0027373

Janssen, O. (2000). Job demands, perceptions of effort-reward fairness and innovative work behaviour. *Journal of Occupational and Organizational Psychology, 73*, 287-302. doi: 10.1348/096317900167038

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*(2), 96-106. doi: 10.1016/j.tsc.2007.04.002

Milgram, R. M., & Milgram, N. A. (1976). Creative thinking and creative performance in Israeli students. *Journal of Educational Psychology, 68*(3), 255-259. doi: 10.1037//0022-0663.68.3.255

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. *The Journal of Creative Behavior, 45*(1), 1-22. doi: 10.1002/j.2162-6057.2011.tb01081.x

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raven, J. (2000). The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48. doi: 10.1006/cogp.1999.0735

Runco, M. A., Abdulla, A. M., Paek, S. H., Al-Jasim, F. A., & Alsuwaidi, H. N. (2016). Which test of divergent thinking is best? *Creativity. Theories – Research - Applications, 3*(1), 4-18. doi: 10.1515/ctra-2016-0001

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*(1), 92-96. doi: 10.1080/10400419.2012.650092

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity, 6*(1), 24-30. doi: 10.1016/j.tsc.2010.06.001

Silvia, P. J., Martin, C., & Nusbaum, E. C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity, 4*(2), 79-85. doi: 10.1016/j.tsc.2009.06.005

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., . . . Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, & the Arts, 2*(2), 68-85. doi: 10.1037/1931-3896.2.2.68

Stein, M. I. (1953). Creativity and culture. *Journal of Psychology, 36*, 311-322. doi: 10.1080/00223980.1953.9712897