| | |
|---|---|
| **Final Technical Report** | Automated AUT scoring using a Big Data variant of the Consensual Assessment Technique |
| **Project** | The development of a valid and usable creativity test requires big-data and psychometrics |
| **Research Grant** | Abbas Foundation Test Development Funds 2016 (Stichting Abbas Fonds Ontwikkelingssubsidie 2016) |
| | website http://www.abbas-fonds.nl |
| **Grant holders** | Claire Stevenson, Matthijs Baas & Han van der Maas |
| **Authors** | Claire Stevenson, Iris Smal, Matthijs Baas, Maike Dahrendorf, Raoul Grasman, Charlotte Tanis, Emma Scheurs, Dana Sleiffer & Han van der Maas |
| **Date** | July 9, 2020 |
| **Licensing** | Disseminated under Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/ |
| **Contact information** | dr. Claire Stevenson |
| | e-mail c.e.stevenson@uva.nl |
| | website http://www.modelingcreativity.org |

## Executive Summary

In this final technical report, we present the results of our Abbas Foundation Test Development Funds project "The development of a valid and usable creativity test requires big-data and psychometrics". The project consisted of two phases: (1) creating a large database of responses for the Dutch version of the "Alternative Uses Task" (AUT) and (2) developing an automated scoring algorithm based on the Consensual Assessment Technique and testing its reliability and validity.

The main aim of the project was to establish a reliable and automated way to score the AUT that will make future data coding faster and more cost-efficient. Meanwhile, the problem of sample-specific scoring would be solved because automated scoring guarantees consistency, i.e. that the same AUT response receives the same creativity score regardless of where the data was collected and scored.

We developed an algorithm that essentially scores AUT responses using the Consensual Assessment Technique based on expert ratings of similar responses from our database of over 70,500 AUT responses. Based on two validation studies, the results show that our algorithm was the best 'rater' and reliably scores new AUT responses similarly to experts. Furthermore, the test-retest and alternate form reliability as well as convergent, discriminant and predicted validity of automated scoring is on par with that of expert scoring. There is still room for improvement, but the current version of our AUT scoring algorithm is a reliable alternative to the time-intensive and costly expert scoring methods.

## Phase 1: Creating a database of Dutch AUT responses

In the first phase, we created a large database which combined AUT data from various creativity researchers in the Netherlands. Our intention was to have a large database of

Dutch AUT data expert ratings that we could use to train an automated scoring algorithm on.

### 1.1. Database Design

The following information (when available) was included for each dataset we received:

- *Research*: principal investigator, institute where research was conducted, year of data collection, design / experimental manipulations, location of research study (online, lab, school, etc.).
- *Rater*: age, experience, role (e.g., research assistant, PhD student, PI), inter-rater reliability.
- *Respondents*: age group, gender.
- *Task*: object (e.g., Brick, Newspaper), task duration, instructions, scoring protocol.
- *Responses*: response, standardized response (spelling mistakes corrected, formulation made uniform), category of response (e.g., for Brick this could be: a building block, art, weight, etc.), time at which response was recorded and/or sequence number of response within respondent.
- *Scores*: ratings for creativity, originality and/or utility (usually a 5-point Likert Scale).

### 1.2. Data Overview

*Table 1. Overview of number of participants and responses currently in our AUT database.*

| object | datasets | respondents | responses |
|---|---|---|---|
| brick | 15 | 2342 | 23621 |
| can | 8 | 1824 | 14957 |
| chord | 7 | 1537 | 12155 |
| fork | 4 | 763 | 5939 |
| paperclip | 6 | 694 | 5298 |
| towel | 5 | 356 | 3163 |
| bottle | 2 | 151 | 1518 |
| book | 3 | 114 | 946 |
| belt | 3 | 115 | 763 |
| newspaper | 2 | 81 | 763 |
| box | 2 | 80 | 686 |
| stick | 2 | 69 | 502 |
| tin | 1 | 34 | 245 |
| **total** | **45** | **8160** | **70556** |

### 1.3. Combined Creativity Score

Most of the data in our database rated creativity on a 5-point Likert scale ranging from (1) not creative at all to (5) highly creative. In our own lab, which provided a substantial portion of the data, we scored the two components of creativity -originality and utility-

rather than a composite creativity score. However, we needed all data to have the same outcome variable. In order to transform our ratings of originality and utility to the more common metric of creativity we needed to combine the two components. We took the following steps. First, we randomly selected 1000 responses per object that had initially been scored only on utility and originality. These were then scored again by two independent raters on the more common 1-5 scale (not creative to highly creative). Then, we investigated which weighted average of originality and utility scores best described the composite creativity score. A set of 11 different models were designed that gave different weights ranging from 0 to 1 at .1 increments for each component (e.g., originality weight .6 and utility weight of .4, or .7 and .3 respectively). Each model was regressed onto the new creativity rating to establish which weighting was best. The combination of 90% originality and 10% utility scores best predicted the composite creativity score. Finally, we used this model to recode all responses that had only utility and originality scores to a composite creativity score.

**Phase 2: Automated AUT scoring**

In the second phase, we developed an automated scoring algorithm to predict the creativity score of new responses to the AUT. To test how well the algorithm performed in comparison to expert ratings, we ran two validation studies. In the first validation study, we gathered data from university students and in the second validation study we collected data from a more generalizable adult sample. We had our algorithm and experts score all of the data and then compared their performance. Reliability (inter-rater, test-retest and alternate form) and validity of our algorithm compared to experts were examined.

This part is structured as follows: (1) description of our automated AUT scoring algorithm, (2) validation study 1 results and (3) validation study 2 results.

### 2.1. AUT scoring algorithm

*Input*
In the table below, you see an example of the input for our algorithm for the object "Brick". You can see that original responses to the AUT were cleaned (e.g., spell-checked, stop words removed, etc.) and this was saved as "cleaned response". The mean of the creativity scores provided by the experts are displayed under the column "creativity rating".

*Table 2. Example input data for AUT brick*

| object | original response | cleaned response | creativity rater 1 |
|---|---|---|---|
| brick | Bankje | bankje | 3.0 |
| brick | Barbecue | barbecue | 3.1 |
| brick | om op te bbq'en | barbecueën | 3.1 |
| brick | Bed | bed | 3.1 |
| brick | Een bed van bouwen | bed bouwen | 2.2 |

*Identifying similar responses*

Previous research suggests that responses to divergent thinking tasks, such as the AUT, group (cluster) together based on how semantically similar the responses are (Acar & Runco, 2019; Beaty & Johnson, 2020; Hass, 2017; Olteţeanu & Falomir, 2016). For example, the responses "build a house" "build a street" and "placing bricks to construct a street" are essentially the same response and semantically very similar. Moreover, they should also receive the same creativity score, as they both suggest using a brick as a building block for streets.

Our algorithm needed some understanding of language to determine the semantic similarity between responses. This can be taught using sentence embeddings, i.e. vector (numeric) representations of sentences that maintain semantic information. Thus, the previous example responses "build a house" and "build a street" should have a very similar sentence embedding (vector representation).

To extract sentence embeddings for the responses to the AUT task, we first extracted word embeddings using Word2Vec. Word2Vec is a method which processes text by 'vectorising' words. Given enough data (such as a Wikipedia corpus), usage and context, Word2Vec can provide highly accurate guesses of a word's meaning. We extracted word embeddings for each word in our AUT database based on pretrained word embeddings from Word2Vec ("Word2vec," 2020, p. 2). We then computed the sentence embeddings by taking the unweighted average of the word embeddings from a response. Semantic similarity was computed by taking the cosine angle between the two sentence embeddings.

*Dataset including semantic similarity*

A dataset was created which contained the creativity scores of at least two raters, the average rater creativity score and the sentence embeddings. Please note that the raters both within and across datasets were not always consistent. For example, "bed bouwen" which translates to "build a bed" was given a 3.05 by raters in one instance and 2.15 in another instance.

*Table 3. Example input data for AUT brick with semantic similarity values.*

| object | cleaned response | creativity rater 1 | creativity rater 2 | mean expert creativity rating | sentence embedding |
|---|---|---|---|---|---|
| brick | bankje | 3.0 | 2.1 | 2.55 | -0.034360 |
| brick | barbecue | 3.1 | 3.1 | 3.10 | 0.030914 |
| brick | bed | 3.1 | 3.0 | 3.05 | -0.031977 |
| brick | bed bouwen | 2.2 | 2.1 | 2.15 | -0.032372 |
| brick | beeldhouwen | 2.2 | 2.2 | 2.20 | 0.036021 |

*Training the algorithm to compute creativity scores*

The first step was to identify which responses were semantically similar and therefore should receive the same creativity score. We used hierarchial clustering to do so. The number of clusters was a hyperparameter, this meant that we first had to identify the

optimal number of clusters. We compared a number of hierarchial clusterings with different numbers of clusters. Then we selected the optimal number of clusters based on the lowest average variance of the creativity rating within all clusters. Our assumption was that low creativity score variance within clusters suggests that the clusters are good at describing the same level of creativity.

The next step was to select an "ideal" response to represent each cluster. We used the sentence embedding that was geometrically most central (remember these are vectors) to the other sentence embeddings in the cluster. This sentence embedding can be considered semantically most similar to all the other sentence embeddings within that cluster.

After this we computed the creativity score to represent all responses in each cluster. This was the average of all of the expert ratings for all of the responses within each cluster. This resulted in the data frame in Table 4.

*Table 4. Example data for AUT brick with clusters and representative responses.*

| cluster | mean expert creativity rating | representative response | sentence embedding |
|---|---|---|---|
| 1 | 2.21 | afbakening | -0.0568580 |
| 2 | 2.92 | anker boot | 0.0348450 |
| 3 | 2.80 | Auto | 0.0332350 |
| 4 | 3.02 | ei bakken | 0.0479155 |
| 5 | 2.58 | Werpen | -0.0203380 |
| 6 | 2.08 | neerleggen zodat iets niet wegwaait | 0.0320292 |
| 7 | 2.60 | voetbal waar bakstenen schoppen verdedigen | -0.0197332 |
| 8 | 2.93 | Basketball | 0.0122180 |
| 9 | 2.15 | bankje stoel | -0.0377145 |

*How the algorithm predicts creativity scores for new responses*
The algorithm computes the creativity of a new response using three steps. First, for every new response, it creates sentence embedding. Second, it computes the semantic distance between this new sentence embedding and each cluster's representative sentence embedding. Third, the new response is assigned the creativity score of the cluster it is semantically most similar to. For example, the new response "bus" would be semantically most similar to the response "car (auto)" in the data table above. Therefore, this new response "bus" would be assigned a creativity score of 2.80, which belongs to the cluster of "car (auto)".

*Testing the algorithm*
Our algorithm was trained on 75% of the AUT database constructed in Phase 1 for the object "Brick" and 80% for the object "Fork". The remaining 20-25% was used to test the

algorithm. The mean absolute percentage error regression loss (where lower values are better) for brick was .25 and for fork this was .20.

### 2.3. Reliability and validity measures

*Inter-rater Reliability*
We wanted our algorithm to be the 'best' expert rater of creativity. Therefore, we looked at the inter-rater reliability (ICC) between the algorithm and the different experts at the response level and the correlation between algorithm and expert scores at the person level (mean creativity rating by each rater). If our algorithm is indeed the 'best' rater then it should have higher ICCs and correlations with each of the other raters than the raters have among themselves.

*Test-retest and Alternate Form Reliability*
Given that we take out the human inconsistencies in scoring, we expected the test-retest reliability and the alternate form reliability for the algorithm to be greater than that of experts.

*Convergent Validity*
We administered the Just Suppose (JS) task and the Product Improvement (PI) task to assess convergent validity. The JS and PI task are both verbal divergent thinking subtests of the TTCT. We expected small to moderate associations between the AUT creativity scores and the JS and PI originality scores. We expected the correlations between the algorithm and the convergent validity measures to be of similar magnitude to that of expert ratings.

*Discriminant Validity*
The Remote Associations Task (RAT) and the Raven intelligence test were used to test the discriminant validity of the AUT. The RAT measures convergent thinking instead of divergent thinking so we expected little to no association between the RAT scores and AUT creativity scores. Similarly, the Raven intelligence test measures abstract reasoning and not divergent thinking so we expected little to no association between the RAT scores and AUT creativity scores. We expected the correlations between the algorithm and the discriminant validity measures to be of similar magnitude to that of expert ratings.

*Predictive Validity*
Self-reported creativity measures such as the Creative Achievement Questionnaire (CAQ, Carson et al., 2005) and the Kaufman Domains of Creativity Scale (KDOCS, Kaufman, 2012) aim to measure general creative ability. We expected small associations between AUT creativity scores and these self-report measures of creativity. Again, we expected the correlations between the algorithm and the predictive validity measures to be of similar magnitude to that of expert ratings.

## 2.4. Validation Studies

### 2.4.1. Validation Study 1 - Student Sample

*Participants*

The sample consists of 110 Dutch first-year psychology students from the University of Amsterdam (69% female; age range 18.3–53.0, M=21.2, SD=3.1 years).

*Materials*

Alternative Uses Task (AUT)
The Alternative Uses task is a common divergent thinking measure used in one form or another in all of the classic divergent thinking test batteries (e.g., Guilford, 1967; Torrance, 1962; Wallach & Kogan, 1965). Participants were asked to name as many creative uses they could think of for an object; in this case, a brick. They had two minutes for this task. Each valid response was judged on a 5-point-scale by two expert raters and also received a score from 1 to 5 by our algorithm. A person's creativity score on the AUT task as a whole was the average of the response ratings; again this was available for each of two raters and the algorithm.

Just Suppose (JS)
The Just Suppose task is part of the Torrance Test of Creative Thinking (Torrance, 2008). In this task, the participant was given a picture and a description of an improbable situation (in this case that the world was so misty that you could only see people's feet) and asked to name as many consequences of this situation as possible. The originality of responses was rated by two experts. The average between experts was used for the response originality score and the mean of all response scores was used as the originality score at the participant level.

Product Improvement (PI)
The Product Improvement task is also part of the TCTT (Torrance, 2008). In this task, participants were presented with a picture and description of a toy monkey and had to come up with as many ways as possible to make the monkey more fun to play with. The average between experts was used for the response originality score and the mean of all response scores was used as the originality score at the participant level.

Raven
The Raven Advanced Progressive Matrices (Raven & Raven, 2003) is a non-verbal abstract reasoning test often used to measure fluid intelligence. The total number of correct responses was used.

RAT
For the Remote Associates Test (Mednick, 1962), participants have to come up with a word that links three seemingly unrelated words together. This version contained six items. The total number of correct responses was used.

CAQ

The Creative Achievement Questionnaire (CAQ, Carson et al., 2005) is a self-report questionnaire about a person's achievements within twelve domains of creativity, such as writing, science and humor.

*Procedure*

The tasks administered for this study formed part of a larger testing battery administered to first year psychology students. The tasks for this project were administered across two testing sessions, both lasting 45 minutes and with one month between the sessions. All tasks were administered by computer in a large exam room.

*Results*

*Data Cleaning*

Initially there were 297 participants that performed the AUT with 2546 responses in total. Ten participants were excluded from the AUT because they did not meet the requirement of providing a minimum of 60% valid responses, a total of 79 responses were removed. Following this, all remaining 58 invalid responses to the AUT and 12 duplicate responses were removed from the dataset. Because we wanted to compare algorithm and expert ratings, 145 responses that the algorithm could not rate because the sentence embedding could not be computed were also removed from the dataset. This resulted in a dataset of 286 participants and 2397 responses. However, not all these participants performed all the other tasks in the test battery, thus only the participants that completed all the tasks used in this validation study were included in the analyses. The final AUT dataset contained 110 participants and 911 responses.

*Inter-rater reliability*

We compared the algorithm and the expert ratings on inter-rater reliability at both the participant level and the response level. For the participant level we looked at the Pearson's correlation between the mean creativity scores by the algorithm versus that of expert raters: $r(108) = .70$, $p < .001$. The correlation between the two experts was: $r(108) = .84$, $p < .001$.

To examine how the algorithm performs on the response level, we examined at the inter-rater reliabilities using the Intraclass Correlation Coefficient (ICC) between the raters and the algorithm, where higher values indicate greater reliability. Table 5 shows that the ICC between the algorithm and the mean expert rating is similar to that of the two experts.

Table 5 also shows the ICCs between experts and the algorithm trained on only on the data from the Modeling Creativity Lab. We see that training with Modeling Creativity Lab data outperforms the algorithm trained on all the data currently in the database; it even outperforms the ICC between the two expert raters.

*Table 5. Study 1 inter-rater reliabilities.*

| Rater 1 | Rater 2 | ICC (95% CI) |
|---|---|---|
| expert 1 | expert 2 | .67 (.64, .71) |
| expert 1 | algorithm (all data) | .52 (.47, .56) |
| expert 2 | algorithm (all data) | .60 (.56, .64) |
| expert 1 | algorithm (MCL data) | .68 (.64, .71) |
| expert 2 | algorithm (MCL data) | .66 (.63, .70) |
| expert mean | algorithm (all data) | .63 (.59, .67) |
| expert mean | algorithm (MCL data) | .75 (.72, .78) |

Figure 1 shows the correlation between the expert scores and the algorithm scores. You can see there is a difference in the spread of the ratings given by the experts and both the algorithms. Ideally, the spread would be more similar. The correlation between experts and algorithm is greater, and therefore better, for the algorithm trained on data only from the Modeling Creativity Lab.

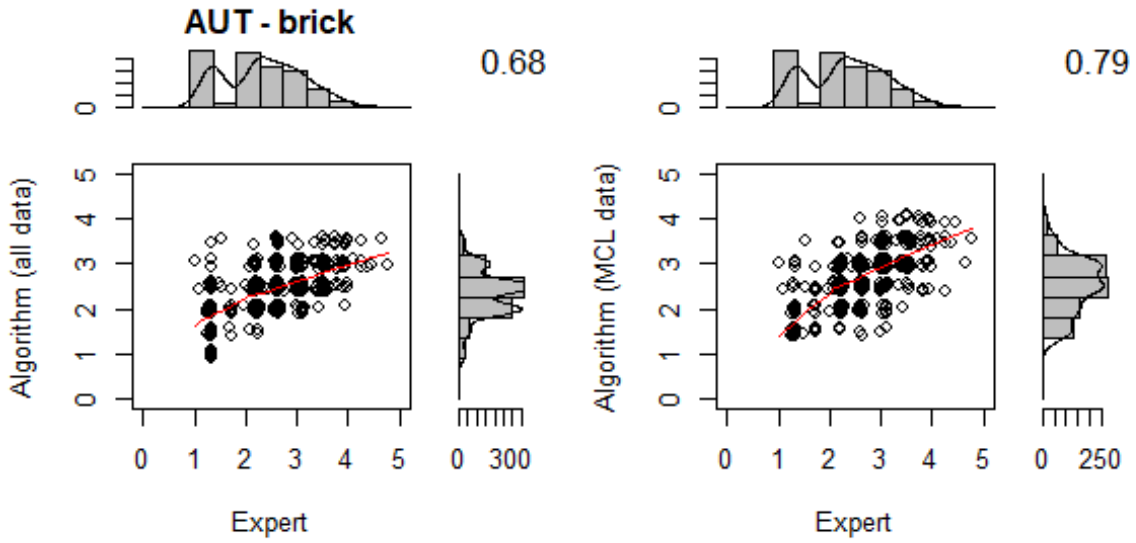


*Figure 1*. Histograms and scatterplots showcasing the spread of the expert scores versus the algorithm scores on the response level.

*Convergent Validity*

We compared the convergent validity of the AUT scoring methods with two other divergent thinking tasks from the TCTT by examining the correlations between the AUT and the PI (*M* = 5.45, *SD* = 3.05) and JS task (*M* = 5.78, *SD* = 3.61). We looked at both the expert (*M* = 2.35, *SD* = 0.37) and algorithm ratings (*M* = 2.32, *SD* = 0.22), where we wanted the algorithm to perform similarly to the experts or better. Table 6 shows that performance on the PI task and the AUT are unrelated, regardless of whether this was expert ratings or the algorithm. We expected a small to moderate correlation here. But, more importantly, we wanted the

algorithm to be on par with experts and this was the case. The correlation between the AUT and JS task performance was small, as expected. The algorithm is nearly on par with experts.

*Table 6. Study 1 convergent validity: correlation between TCTT subtests and AUT performance.*

| | **expert ratings** | **algorithm ratings** (all data) |
|---|---|---|
| PI originality | .02 (-.16, .21) | -.01 (-.20, .18) |
| JS originality | .25 (.07, .42) | .14 (-.05, .32) |

*Discriminant Validity*

The average score on the RAT was 2.39 (*SD* = 1.36) and the average score on the Raven was 10.72 (*SD* = 4.51). The correlations between the AUT, for both the expert and algorithm scores, and the RAT and Raven are low, as was expected. The discriminant validity looks good based on the RAT and the Raven and the algorithm is on par with experts.

*Table 7. Study 1 discriminant validity: correlation between RAT, Raven and AUT performance.*

| | **expert ratings** | **algorithm ratings** (all data) |
|---|---|---|
| RAT | .09 (-.10, .27) | .01 (-.17, .20) |
| Raven | .12 (-.07, .30) | .09 (-.10, .27) |

*Predictive Validity*

The CAQ was used to test the predictive validity of the AUT. Participants on average scored 6.15 (*SD* = 4.32) on the CAQ. We expected small correlations between the CAQ and the AUT. In Table 8 we can see that there actually is no correlation. The algorithm does perform similarly to the experts.

*Table 8. Study 1 predictive validity: correlation between CAQ and AUT performance.*

| | **expert ratings** | **algorithm ratings** (all data) |
|---|---|---|
| CAQ | .06 (-.13, .24) | .01 (-.17, .20) |

*Reliability and Validity Results using algorithm trained solely on Modeling Creativity Lab data*

In Table 9 the same correlations are shown, but here the algorithm was trained using data scored by the Modeling Creativity Lab. The algorithm performs better when only using this data, even outperforming the experts in terms of discriminant validity assessed with the Raven. This makes sense because the AUT for the validation studies were scored using the Modeling Creativity Lab protocol.

*Table 9. Study 1 correlations between AUT expert ratings and both algorithm versions with all validity measures.*

| | **expert ratings** | **algorithm ratings** (all data) | **algorithm ratings** (MCL data) |
|---|---|---|---|
| expert ratings | | | |
| algorithm ratings (all data) | .70 (.59, .78) | | |
| algorithm ratings (MCL data) | .80 (.72, .96) | .76 (.66, .83) | |
| PI | .02 (-.16, .21) | -.01 (-.20, .18) | .03 (-.15, .22) |
| JS | .25 (.07, .42) | .14 (-.05, .32) | .23 (.05, .40) |
| CAQ | .06 (-.13, .24) | .01 (-.17, .20) | .04 (-.15, .23) |
| Raven | .12 (-.07, .30) | .09 (-.10, .27) | .01 (-.18, .19) |
| RAT | .09 (-.10, .27) | .01 (-.17, .20) | .05 (-.14, .23) |

### 2.4.2. Validation Study 2 – General Population Sample

*Participants*

For this validation study, 132 participants were recruited at the Amsterdam University of Applied Sciences and through online advertisements. This way, the sample was more diverse than the one used in the first validation study. 116 participants were included in the analyses (see reasons for exclusion in the results section); 37 males and 79 females with ages ranging from 18 to 65+ ($M$ = 24.57, $SD$ = 9.82).

*Instruments*

Alternative Uses Task (AUT)
See description of the task in section 2.4.1. In this study all the participants completed the AUT two times. Half of the participants were presented with the same object two times (brick-brick/fork-fork) so that we could examine the test-retest reliability of the AUT. The other half was presented with two different objects (brick-fork/fork-brick) so that we could examine the alternate-form reliability. The participants were randomly assigned and the groups were about the same size.

KDOCS
The Kaufman Domains of Creativity Scale (KDOCS, Kaufman, 2012)is a self-report questionnaire that looks at common perceptions of creativity in five domains: Self/Everyday, Scholarly, Performance, Mechanical/Scientific, and Artistic.

Creative Achievement Questionnaire
See description in section 2.4.1.

RAT
See description in section 2.4.1.

*Procedure*

Participants that expressed interest in participating in the study were provided with a link to the test battery. They could choose if they wanted participate in their own time or in a supervised quiet classroom at the Amsterdam University of Applied Sciences. The order of the tasks was the same for all the participants, except for the order of the two AUT items (see description Instruments). First AUT item 1 was administered, then they filled in the K-DOCS, the CAQ, and the RAT. Last they completed the AUT item 2. The test battery, including providing consent and reading instructions, lasted no longer than one hour. Participants received 10 euros compensation for their participation.

*Results*

*Data Cleaning*
Initially there were 132 participants and 2312 responses. 16 participants did not meet the requirement of a minimum of 60% valid responses on both AUT's to be included, which resulted in excluding 416 responses. Following this, the remaining 110 invalid responses on

the AUT were removed from the dataset. Again, because we are looking at the performance of the algorithm ratings as well as the expert ratings, 99 responses that the algorithm could not rate were also removed from the datasets. This resulted in a dataset containing 116 participants and 1687 responses.

*Inter-rater Reliability*

The reliability of the algorithm ratings compared to expert ratings on the person level was assessed using the correlation between the algorithm and the mean of the expert raters, $r(114) = .80$, $p < .001$. This is very similar to the correlation between both the experts, $r(114) = .81$, $p < .001$. The algorithm performs well in this regard.

The inter-rater reliability at the response level was examined using the Intraclass Correlation Coefficient (ICC). The invalid responses where both experts scored zero and where the algorithm was not able to make predictions, were removed prior to analyses, in 939 responses for "brick" and 748 responses for "fork".

Table 10 shows three expert raters. Rater 1 and rater 2 are expert raters from the Modeling Creativity Lab (MCL), rater 3 is an expert rater from Baas's lab. Rater 3 was included here to better understand why the algorithm trained only on the Modeling Creativity Lab's data generally performed better than the algorithm trained on all the data. Next to the Modeling Creativity Lab, Baas's Lab provided a significant portion of the data in the database. In his lab, the AUT is scored on originality, instead of on originality and utility as done by the Modeling Creativity Lab. This results in quite different ratings between labs, as can be seen in Table 10. For the main analyses, the expert ratings of rater 1 and 2 are combined.

*Table 10. Study 2 Inter-rater reliabilities between experts and algorithms.*

| **Rater 1** | **Rater 2** | **Brick ICC (95% CI)** | **Fork ICC (95% CI)** |
|---|---|---|---|
| expert 1 | expert 2 | .61 (.57, .65) | .69 (.65, .72) |
| expert 1 | expert 3 | .01 (-.05, .08) | .01 (-.07, .08) |
| expert 2 | expert 3 | -.10 (-.16, -.04) | -.09 (-.16, -.01) |
| expert 1 | algorithm (all data) | .53 (.48, .58) | .71 (.67, .74) |
| expert 2 | algorithm (all data) | .52 ( .47, .56) | .63 (.58, .67) |
| expert 3 | algorithm (all data) | -.07 (-.14, -.01) | -.08 (-.15, 0) |
| expert 1 | algorithm (MCL data) | .59 (.55, .63) | .70 (.67, .74) |
| expert 2 | algorithm (MCL data) | .63 (.59, .67) | .67 (.62, .70) |
| expert 3 | algorithm (MCL data) | -.12 (-.18, -.06) | -.13 (-.20, -.06) |
| expert 1 + 2 | algorithm (all data) | .61 (.57, .65) | .74 (.70, .77) |
| expert 1 + 2 | algorithm (MCL data) | .70 (.67, .73) | .76 (.72, .79) |

The ICC's between rater 1 and rater 2, the expert mean, and the algorithm trained on either all data or only Modeling Creativity Lab data, for both "brick" and "fork", all meet the minimum requirement of .6 or higher. We wanted the ICC of the between expert ratings and the algorithm to be higher than the ICC between rater 1 and rater 2. We can see that for "brick" the algorithm trained on all the data performs on par with experts. The algorithm

trained on the Modeling Creativity Lab's data outperforms experts. For "fork" both versions of the algorithm outperform experts. 1 – 2

Figure 2 shows the spread of the ratings given by the experts and both of the algorithms on the response level. Especially for "brick", the algorithm appears to "normalize" the distribution of the ratings; the algorithm based on all the training data does this even more than the algorithm based on the Modeling Creativity Lab's training data.

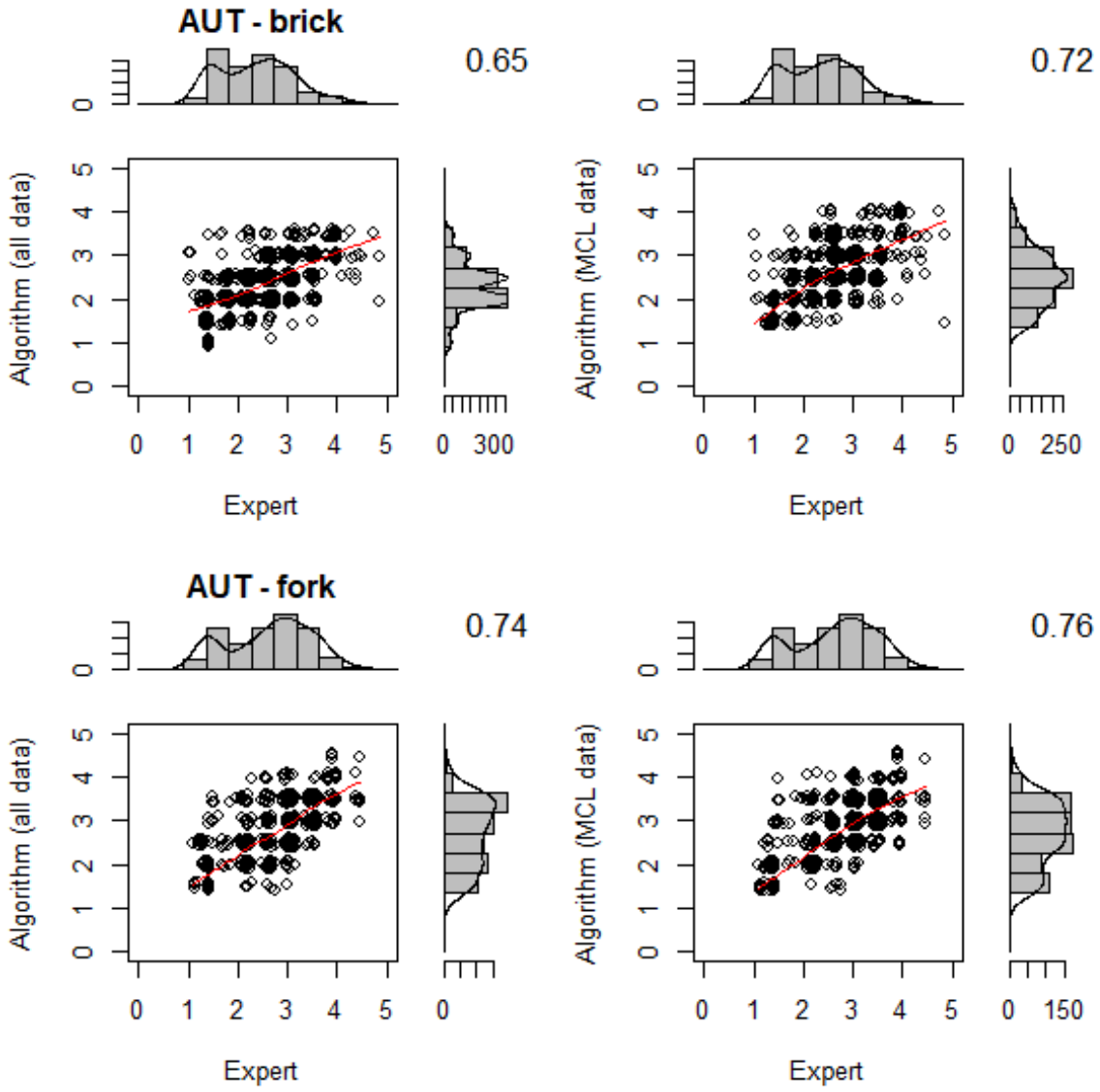


*Figure 2.* Histograms and scatterplots showcasing the spread of the expert scores versus the algorithm scores on the response level for both brick and fork.

The correlations show that the algorithm performs better for "fork" than "brick" data and that the correlations between experts and the algorithm are higher for data trained on Modeling Creativity Lab data rather than all data.

*Test-retest reliability*

To set a benchmark correlation for test-retest and alternate form reliability of what we could realistically hope for, we looked at correlation between the fluency scores between the first and second AUT: $r(114) = .57$, $p < .001$.

Table 11 shows the test-retest reliability. The correlations between the AUT scores on the first and second administration of the AUT for the object "fork" is very high, with both the algorithms outperforming the experts and all the correlations exceeding the previously mentioned benchmark. This is not true for the test-retest reliability for "brick". Experts perform slightly better than both algorithm versions and all correlations are similar in magnitude to the benchmark of .57.

*Table 11. Study 2 test-retest reliability*

| | **fork-fork** | **brick-brick** |
|---|---|---|
| Expert ratings | .70 (.43, .85) | .61 (.34, .78) |
| Algorithm ratings (all data) | .73 (.48, .87) | .55 (.25, .75) |
| Algorithm ratings (MCL data) | .80 (.60, .90) | .49 (.18, .71) |

*Alternate form reliability*

The alternate-form reliability is shown in Table 12. Here we can see that the correlations are lower than for the test-retest reliability, which is to be expected because now we are looking at two different items rather than repeated administration of the same item. Experts outperform the algorithm for both fork-brick and brick-fork. On fork-brick the algorithm trained with all data is more reliable across items than the algorithm trained with only Modeling Creativity Lab data. However, the correlations for brick-fork are problematic; the correlation between the expert scores is surprisingly low and the algorithms even show negative correlations across administrations.

*Table 12. Study 2 alternate form reliability*

| | **fork-brick** | **brick-fork** |
|---|---|---|
| Expert ratings | .42 (.05, .68) | .16 (-.23, .51) |
| Algorithm ratings (all data) | .35 (-.02, .64) | -.28 (-.60, .11) |
| Algorithm ratings (MCL data) | .29 (-.10, .60) | -.19 (-.53, .21) |

*Discriminant Validity*

To assess the discriminant validity, the AUT, a divergent thinking task, was compared to the RAT, a convergent thinking task. We did this for both the algorithm scores and the expert scores: $r(114) = .22$, $p = .02$; $r(114) = .30$, $p < .001$, respectively. The correlations are slightly higher than expected, however, the algorithm and the experts do perform similarly in regard to discriminant validity; the algorithm actually outperforms the experts.

*Predictive Validity*

The CAQ and the KDOCS were used to assess the predictive validity of the AUT. We expected small correlations between these tests and the AUT. The correlations between the CAQ and the AUT scores, for the algorithm and expert raters, are $r(114) = .09$, $p = .32$ and $r(114) =$

.16, $p$ = .08. The correlation between the K-DOCS and the AUT is $r(114)$ = .07, $p$ = .48 for the algorithm, and $r(114)$ = .01, $p$ = .95 for the experts. We can see that the expert raters performed better with regard to predictive validity for the CAQ, but worse for the K-DOCS. All of these correlations were lower than expected.

*Table 13. Study 2 means and standard deviations of the AUT scores and validity measures.*

| | M | SD |
|---|---|---|
| AUT expert ratings | 2.51 | 0.38 |
| AUT algorithm ratings (all data) | 2.48 | 0.31 |
| AUT fluency | 14.58 | 5.68 |
| KDOCS | 154.15 | 20.06 |
| CAQ | 3.47 | 3.01 |
| VF | 14.82 | 7.39 |
| RAT | 8.18 | 4.72 |

In the previous validation study, we also looked at how the algorithm performed when trained only on data from the Modeling Creativity Lab. Table 14 shows the correlations between all administered tasks and both versions of the algorithm. The differences are small and the Modeling Creativity Lab trained algorithm does not perform better than the algorithm trained on all data algorithm with regard to validity.

*Table 14. Study 2 correlations between validity measures and the AUT scored by experts, algorithm trained on all data and algorithm trained on Modeling Creativity Lab (MCL) data.*

| | AUT expert ratings | AUT algorithm ratings (all data) | AUT algorithm ratings (MCL data) |
|---|---|---|---|
| AUT algorithm ratings (all data) | .80 (.72, .86) | | |
| AUT algorithm ratings (MCL data) | .86 (.80, .90) | .87 (.82, .91) | |
| KDOCS | .01 (-.18, .19) | .07 (-.12, .25) | -.02 (-.20, .17) |
| CAQ | .16 (-.02, .34) | .09 (-.09, .27) | .10 (-.08, .28) |
| RAT | .30 (.13, .46) | .22 (.04, .39) | .26 (.08, .42) |

## Conclusion & Discussion

We developed an algorithm that essentially scores AUT responses based on expert ratings of similar responses from our database of >70,500 AUT responses. Our results show that our algorithm was the best 'rater' and was on par with experts for nearly all measures of reliability and validity across both studies. Conclusions based on both studies will now be discussed for each of the reliability and validity measures.

*Inter-rater reliability*

The inter-rater reliability scores were highest when one of the raters was the algorithm (rather than an expert), which means we can consider our algorithm the 'best' expert rater. This was especially the case when our algorithm was trained with data scored according to the Modeling Creativity Lab (MCL) protocol rather than all of the data in our AUT database. This makes sense because MCL experts also scored the data for both validation studies, so the scoring by the algorithm and the experts was essentially based on the same protocol.

*Test-retest reliability*

In study 2, we assessed the test-retest reliability of the AUT when scored by experts and when scored by the two versions of the algorithm. Test-retest reliability was highest for the MCL trained algorithm for the object "fork" ($r$=.80 for MCL algorithm, $r$=.70 for experts) and it was highest for the experts for the "brick" object ($r$=.61 for experts, $r$=.49-.55 for algorithm). In both cases, there were less than 30 participants and the confidence intervals were large and overlapping. Thus, we can conclude that the algorithm and expert scored AUTs had similar test-retest reliability.

*Alternate form reliability*

In study 2, we also assessed the alternate form reliability of the AUT when scored by experts and when scored by the two versions of the algorithm. When participants were first administered the AUT fork and then the AUT brick the algorithm alternate form reliability was lower than that of experts ($r$=.29-.35 versus $r$=.42 respectively), but of similar magnitude. When participants first solved the AUT brick and then AUT fork the correlation was surprisingly low for experts and correlations were negative for the algorithms. It appears that quite a few of the participants who started with the AUT brick misunderstood the task and therefore these correlations are based on a small number of participants ($N$=27), which explains the large overlapping confidence intervals and may explain the surprising results. We can conclude that the algorithm and expert scored AUTs had similar alternate form reliability.

*Convergent validity*

In study 1, we examined convergent validity by comparing people's AUT creativity scores to their originality scores on two TTCT subtests, the Just Suppose task and the Product Improvement task. Just Suppose task performance was related to AUT performance, Product Improvement performance was not. In both cases, algorithm performance was on par with expert performance. This was especially the case for the algorithm trained with data scored according to the Modeling Creativity Lab protocol. This makes sense because experts from our lab also scored the TTCT tasks.

*Discriminant Validity*

To assess the discriminant validity, the AUT, a divergent thinking task, was compared to the RAT, a convergent thinking task, and the Raven, a fluid intelligence test. In study 1, the correlations between the AUT, for both the expert and algorithm scores, and the RAT and Raven were low, as was expected and the algorithm was on par with experts. In study 2,

only the RAT was administered. For both experts and algorithm these correlations were unexpectedly high (range .22 - .30), but of similar magnitude, so again the algorithm performed similarly to experts.

*Predictive Validity*

The CAQ and the K-DOCS were used to assess the predictive validity of the AUT. We expected small correlations between the results of the two self-rating scales and the AUT. In study 1, the algorithm performed similarly to the experts, although the correlation between the CAQ and AUT was nearly zero (for both experts and algorithms). In study 2, the predictive validity of experts (r=.16) was better than the algorithm (r=.09) for the CAQ, but worse for the K-DOCS (algorithm r=.07, experts r=.00). However, all of these correlations were lower than expected and did not differ substantially in magnitude. Thus, we conclude that in exception to the CAQ in study 2, the algorithm and expert scored AUT had similar predictive validity.

*Limitations & Future Directions*

Our algorithm scored new AUT responses by finding semantically similar responses in our AUT database and then giving them the average expert rating. Therefore, the most important features our algorithm used to score the AUT were the expert ratings and the sematic similarity between responses. However, a number of other features could help improve automated AUT scoring. For example, how frequently a response is given, how unique a response is (inverse of frequency), and how semantically similar a response is to the object in question (Tsai, 2020). We plan to include these and similar features in the next version of our automated AUT scoring algorithm.

Furthermore, given the high reliance of our algorithm on the semantic similarity of AUT responses, it would be useful to explore the different ways to obtain word and phrase vectors to compute semantic distances between responses. We chose Word2Vec early on in 2016 this project. However, more recent alternatives such as FastText, GloVE, BERT and ELMo are also promising (Scheurs, 2020; Tsai, 2020).

There are a number of machine learning approaches to find the optimal combination of features to predict outcomes, in this case ratings for new AUT responses. Our approach was based on the Consensual Assessment Method using the average of expert ratings. But, we could combine expert ratings by weighting them and also add additional features (e.g., such as frequency as mentioned above). In future work, we recommend comparing a number of different machine learning methods –from ridge regression to random forests- to find the best possible combination of features to automatically rate AUT responses (e.g., Tsai, 2020).

Currently our database and algorithm were limited to the Dutch language. Given the simplicity of automated translation and the vast natural language processing possibilities for English language text we have three suggestions for future versions. First, the database can be extended with data from all languages that can be automatically translated to English. Second, features based on semantic similarity between responses could be measured both within each language as well as across all languages providing a rich set of features for the algorithm to base its automated scoring on. Third, the database itself could

be an excellent source of research data for creativity researchers interested in cross-language investigations of how people solve the AUT.

In addition, our algorithm and the validation study tests were limited to the "brick" and "fork" objects. We trained the algorithm separately for each object; however, pilot studies have shown that our algorithm can be generalized to all objects by including semantic distance measures not only between responses from the same object, but also with responses to other objects and the object itself (Sleiffer, in prep; Tsai, 2020). This is the direction we recommend for future automated scoring algorithms, especially once more data becomes available from other language sources.

Finally, we noticed that labs can differ substantially in how their experts rate creativity. This could be because there is still some debate on what constitutes creativity and whether and how it should be scored with divergent thinking tasks (Reiter-Palmon et al., 2019; Runco & Jaeger, 2012; Simonton, 2018; Stevenson et al., 2020). We think that researchers who use our automated scoring algorithms should be able to choose how they want to score the AUT. Therefore, in the future we recommend creating different algorithms to score creativity, originality, and/or utility according to different lab's protocols.

## References


Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 153–158. https://doi.org/10.1037/aca0000231

Beaty, R., & Johnson, D. R. (2020). *Automating Creativity Assessment with SemDis: An Open Platform for Computing Semantic Distance* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nwvps

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. https://doi.org/10.1207/s15326934crj1701_4

Guilford, J. P. (1967). Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 298–308. https://doi.org/10.1037/a0029751

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Olteţeanu, A.-M., & Falomir, Z. (2016). Object replacement and object composition in a creative cognitive system. Towards a computational solver of the Alternative Uses Test. *Cognitive Systems Research*, *39*, 15–32. https://doi.org/10.1016/j.cogsys.2015.12.011

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In *Handbook of nonverbal assessment* (pp. 223–237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Scheurs, E. (2020). *Automated Categorization of the Alternative Uses Task: A Neural Network Approach* [Master's thesis, University of Amsterdam]. http://modelingcreativity.org/blog/wp-content/uploads/2020/07/Emma-Scheurs-Final-Version-Master-Thesis.pdf

Simonton, D. K. (2018). Defining Creativity: Don't We Also Need to Define What Is Not Creative? *The Journal of Creative Behavior*, *52*(1), 80–90. https://doi.org/10.1002/jocb.137

Sleiffer, D. (in prep). *Automated AUT scoring based on Semantics and the Consensual Assessment Technique* [Master's internship report]. University of Amsterdam.

Stevenson, C., Baas, M., & van der Maas, H. (2020). *A Minimal Theory of Creative Ability* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/t5e3g

Torrance, E. Paul. (1962). *Guiding creative talent*. Prentice-Hall, Inc. https://doi.org/10.1037/13134-000

Torrance, E. Paul. (2008). *The Torrance Tests of Creative Thinking—Norms—Technical manual—Figural (streamlined) forms A and B.* Scholastic Testing Service.

Tsai, Y. (2020). *Semantic-Based Algorithm for Scoring the Alternative Uses Tests* [Master's thesis, University of Amsterdam]. http://modelingcreativity.org/blog/wp-content/uploads/2020/07/Tsai_Y_BDS_Thesis_report_11695986_PML.pdf

Wallach, M. A., & Kogan, N. (1965). A new look at the creativity-intelligence distinction. *Journal of Personality*, *33*(3), 348–369. https://doi.org/10.1111/j.1467-6494.1965.tb01391.x

Word2vec. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Word2vec&oldid=966439837