JOURNAL OF
Creative Behavior
Publication of the Creative Education Foundation™

BENJAMIN GOECKE
PAUL V. DISTEFANO
WOLFGANG ASCHAUER
KURT HAIM
ROGER BEATY
BORIS FORTHMANN

Automated Scoring of Scientific Creativity in German

ABSTRACT

Automated scoring is a current hot topic in creativity research. However, most research has focused on the English language and popular verbal creative thinking tasks, such as the alternate uses task. Therefore, in this study, we present a large language model approach for automated scoring of a scientific creative thinking task that assesses divergent ideation in experimental tasks in the German language. Participants are required to generate alternative explanations for an empirical observation. This work analyzed a total of 13,423 unique responses. To predict human ratings of originality, we used XLM-RoBERTa (Cross-lingual Language Model-RoBERTa), a large, multilingual model. The prediction model was trained on 9,400 responses. Results showed a strong correlation between model predictions and human ratings in a held-out test set (n = 2,682; r = 0.80; CI-95% [0.79, 0.81]). These promising findings underscore the potential of large language models for automated scoring of scientific creative thinking in the German language. We encourage researchers to further investigate automated scoring of other domain-specific creative thinking tasks.

Keywords: creativity, automated scoring, scientific creativity, large language models.

AUTOMATED SCORING OF CREATIVE THINKING TASKS

Attempts of automated scoring of creative thinking tasks are surprisingly old (Paulus, 1970). At the end of the 1960s, a regression-based prediction model of scores for the Torrance Test of Creative Thinking was proposed, which is one of the most widely used tests of divergent thinking. Simple text mining statistics such as the number of words in a response were used to predict scoring by human raters and their approach has provided reasonable predictions of human ratings in recent efforts (Forthmann & Doebler, 2022). Benefits of automated scoring are self-evident: for example, lower associated costs due to less labor, less prone to human biases (although this assumption can be considered contested), availability of full computerized assessment. Application of creativity measures for individual purposes and possibly personnel selection (compare responses against huge reference groups of prior data).

Generally, 15 years ago the idea of automated scoring revived with most works relying on latent semantic analysis or other word vector models of meaning (Bossomaier, Harré, Knittel, & Snyder, 2009; Dumas & Dunbar, 2014; Forster & Dunbar, 2009; Green, Kraemer, Fugelsang, Gray, & Dunbar, 2012). Initially, validity findings were rather inconsistent which is most likely attributable to technical issues such as the known elaboration-bias inherent in latent semantic analysis. Focusing on other word vector models such as GloVe (Dumas, Organisciak, & Doherty, 2021), using multiplicative compositional approaches (Beaty & Johnson, 2021), or maximum semantic distance (Yu et al., 2023) seemed to clearly improve automated scoring quality. Most recent work switched to Large Language Models (LLMs) which can closely mimic the scoring of human raters (Organisciak, Acar, Dumas, & Berthiaume, 2023).

LLMs can score divergent thinking tasks automatically (Organisciak et al., 2023) and already achieve impressive prediction of human ratings without additional training (i.e., at zero-shot). The models' capacity to score divergent thinking tasks improves substantially with small amounts of training data, resulting in

correlations of up to $r \sim .80$ at the response level, which was argued to be the possible ceiling. After all, inter-rater reliabilities between any two or more human scorers rarely do exceed such values as well.

COMPLEMENTING EFFORTS OF SCORING CREATIVITY AUTOMATICALLY

Most studies that use automated scoring techniques rely on data from the English language and a limited set of creativity indicators, primarily the alternate uses task (e.g., Guilford, 1967). However, recent advancements have expanded our ability to automatically score creativity across different languages (Forthmann & Doebler, 2022; Patterson, Merseal, et al., 2003; Zielińska, Organisciak, Dumas, & Karwowski, 2023) and tasks (Acar, Organisciak, & Dumas, 2023; Cropley & Marrone, 2022; Patterson, Barbot, Lloyd-Cox, & Beaty, 2023).

Although some progress has been made in transferring knowledge gained in the English language domain to other languages, such as German (e.g., Forthmann & Doebler, 2022; Patterson, Merseal, et al., 2023), further evidence is needed to evaluate the predictive performance of automated scoring approaches for this language. With roughly an estimated 100–160 million individuals speaking German as their first or second language in the world, German belongs to the 20 most spoken languages in the world (c.f., https://de.statista.com; full link provided in the Supplementary Materials in the OSF). Due to the large potential of automated creativity scoring and the considerable number of individuals on earth not using English as their lingua franca in their daily lives, it makes sense to extend previous efforts of automated creativity scoring to further languages, such as German as well.

Second, today a large arsenal of measurement instruments of creativity is available (Weiss, Wilhelm, & Kyllonen, 2021) and new measurement approaches are continuously developed and evaluated. This includes both new measurement instruments, but also refined constructs that can be measured. For example, measuring domain-specific creative abilities has become increasingly popular over the last few years, and new measurement instruments for domains like emotions (Weiss, Olderbak, & Wilhelm, 2023), music (Merseal et al., 2023), or scientific creative thinking (Aschauer, Haim, & Weber, 2022) were developed and validated. These developments can also be understood as a call for broadening our capabilities of automatically evaluating the creativity of test-takers, as future multivariate studies could benefit from less labor-intensive scoring practices. Hence, testing existing or new algorithms for automatically scoring creativity tasks would benefit from extending the scope of tasks to which such algorithms can be applied.

AIM OF THE CURRENT STUDY

Most previous work regarding automated scoring of creativity has focused mostly on the English language, on the classical Alternate Uses Task, or other verbal divergent thinking tasks such as the Consequences Task. While there exists now psychometric work on automated scoring of creative writing tasks (Johnson et al., 2022) or figural divergent thinking tasks (Cropley & Marrone, 2022; Patterson, Barbot, et al., 2023), work on domain-specific creative thinking tasks in other languages than English is still lacking. Hence, the main aim of this work is to address this gap in the literature by reporting a study that shows that automated scoring using a LLM of a scientific creative thinking task (i.e., a domain-specific creative thinking task) in the German language is feasible. Scientific creative thinking can be conceptualized as a *divergent problem solving ability in science* (DPAS; Aschauer et al., 2022). In this paper, the operationalization focuses on originality (e.g., uncommon, remote, and clever ideas). Given the increasing importance of creativity in science education and research on fostering creative thinking in education, it is necessary to improve both our measurement instruments for assessing this ability and our scoring methods for evaluating test takers' responses. The research objectives of the present study were not pre-registered.

METHOD

SAMPLE AND PROCEDURE

The sample for the current study consisted of $N = 1{,}272$ students from 5th to 12th grade from 52 classes (middle schools; academic-track schools, and vocational-track schools) altogether in Austria. The age range was approximately 11 to 18 (please note that this information was not collected, but that the grade is a good proxy for the age of a student in [please will be disclosed after peer-review]). Approximately, 51% of the students were females. The measurement was part of a larger multivariate study consisting of two measurement time points. As these design-related characteristics of the study are not inherently associated with our research aims, we will refrain from going into more detail here (please see Aschauer et al., 2022 for more comprehensive information).

MEASURES

For the current study, we use one specific item of a test battery designed to measure scientific creativity by DPAS (please see Aschauer et al., 2022 for a comprehensive description of the test battery). The employed item is part of the DPAS subscale for divergent ideation in experimental tasks. Students were asked to come up with creative reasons for the following hypothetical phenomenon: "In 2000, the room temperature of a classroom was measured for 1 year, and the average temperature at that time was 18°C. Today, 19 years later, the mean temperature of this classroom was 22°C. Give as many different reasons as possible why the room temperature in this classroom is 4°C higher today." Students were informed that they were participating in a creativity test that will test the richness of their ideas (e.g., "find many different ideas"). They were instructed to think of as many responses as possible. The test was administered computerized. In total, N = 18,653 responses were generated and considered for analysis in this work.

HUMAN SCORING RATING DESIGN

The rating design for the human scoring was based on a planned missingness design (Forthmann, Goecke, & Beaty, 2023) that we obtained through a simulation-based approach in R (R Core Team, 2022). Specifically, we based our rating design on a simulation based on three available empirical data sets: the data of an Alternate Uses Task (N = 209 participants with n = 3,236 responses; c.f., Patterson, Barbot, et al., 2023; Patterson, Merseal, et al., 2023), and two data sets of a scientific creativity test (N_1 = 1,147 with n_1 = 4,369 responses; N_2 = 146 with n_2 = 7,923 responses; c.f., Beaty et al., under review). We applied a generalized partial credit model (GPCM; Muraki, 1992) for the simulation and tested five competing scenarios in our simulations with different parameters in order to find the best planned missingness design regarding cost-reliability trade-off for our purposes. Please see the Supplementary Materials for more information regarding the data generation process of the simulation.

We found that a design with N = 40 human raters, and 4 ratings per response for 50% of all responses would yield an average reliability of .822 which was deemed appropriate considering the trade-off between costs, human labor and reliability as a good cutoff is considered to be .80 (c.f., the factor determinacy index, Ferrando & Lorenzo-Seva, 2018). Explicitly, this rater design meant that on average, each rater had to rate 1,569 responses (range: 1,551–1,701 responses).

We used the obtained rater design to prepare 40 single rating sheets containing, in total, all the available responses according to our planned missingness rater design. We then recruited 40 human raters with sufficient German skills via Prolific, who rated the German responses. Each rating sheet was assigned to one rater (we provide one example rating sheet including instructions for raters via the OSF). The raters received detailed instructions regarding scoring with the rating sheets and were motivated to avoid missingness. The instructions were available to them at all times. In addition to that, three of the co-authors picked 3 example responses for each possible response category and explained why each response should be rated this way. Raters were reimbursed with 30$ (12$/h), as we estimated that it would take them about 2.5 hours to score the responses assigned to them.

Once the ratings were completed, we tested the data not only based on the initially imposed GPCM but also decided to test competing measurement models, such as the Graded Response Model (GRM; Samejima, 1968). The reliabilities and correlations for both models are depicted in Table 1. The GRM fitted the data slightly better, so we decided to use the factor scores obtained by this model for further evaluation (i.e., human scored "creativity ratings"). Please note, however, that the correlation between both models' parameters amounted to r = .99. These parameters were the basis for the automated scoring approach which will be outlined next. Please note that for validity purposes, we provide correlations between the human ratings (i.e., factor scores) and fluency and flexibility scores of the scientific creativity task in the

TABLE 1. Fit Indices, Empirical Reliabilities, and Correlations of Applied Rater Models

Model	AIC	BIC	$r_{\theta\theta}$	$\sqrt{r_{\theta\theta}}$
GCPM	124348	125895.6	0.731	0.855
GRM	124150	125697.5	0.742	0.861

Note. $r_{\theta\theta}$ = Empirical Reliability; $\sqrt{r_{\theta\theta}}$ = Correlation between estimated latent response scores and the true response scores.

Supplementary Materials (Figure S2). We provide all materials needed to replicate our approach in an open repository: https://osf.io/aw95p/.

AUTOMATED SCORING APPROACH

We fine-tuned a large language model (LLM) to predict the human ratings of the problem-solving task. The LLM we used was XLM-RoBERTa-base (Conneau et al., 2020), which is an open-source pre-trained multi-lingual language model with 125 M parameters based on Meta's RoBERTa model (Liu et al., 2019). This model was trained using a Transformer architecture (Devlin, Chang, Lee, & Toutanova, 2019), as is the case with models of the Bidirectional Encoder Representations from Transformers (BERT) architecture. XLM-RoBERTa specifically was trained in over 100 languages, thus it can predict ratings for text in German.

To identify the model settings (hyperparameters) that elicit the best performance, we engaged in a hyperparameter search using the Optuna Python package (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). Prior to this search, the data were deduplicated, reducing the data set size from 18,314 to 13,423 unique responses. Subsequently, the dataset was randomly split into training, validation, and held-out test data sets, utilizing a 70/10/20 ratio, consistent with the best practices in machine learning (Zhou, 2021). The model input was the task prompt "Ein kreativer Grund für Temperaturveränderungen in einem Klassenzimmer ist" ("A Creative Reason for Temperature Changes in a Classroom is") immediately followed by the participant response. An example of a highly creative response was "Ein übernatürlicher Mensch aus einer anderen Dimension machte die Sonne heißer, so dass es bei uns auch heißer ist" ("A supernatural person from another dimension made the sun hotter so that ours is hotter too.").

Across 230 trials, we searched over three hyperparameters (learning rate, batch size, and the number of training epochs). Learning rate corresponds to the extent to which training episodes influence the model weights, where higher rates result in greater change during each training batch. The range of learning rates we searched over was 5e-07 to 5e-02. Batch size is the number of responses that the model receives feedback on at a time and we tested values of 4, 8, 16, and 32. Lastly, we surveyed an epoch range of 10 to 150 to determine the optimal number of passes through the training dataset. Optuna searches over these hyperparameter settings for the configuration that minimizes the mean square error (MSE) between the provided human ratings and the model predicted-ratings of the validation set. The trial with the best performance on the validation set (n = 1,341) employed a learning rate of 4.37e-06, ran for 11 epochs, and utilized a training batch size of 4. Using these hyperparameters, rating predictions generated by XLM-RoBERTa correlated positively with human ratings (r = 0.81; CI 95% [0.79, 0.83]). The settings of this trial were used to evaluate the held-out test set. We removed outliers in the model predictions that were 3 standard deviations from the mean (please note, however, that we report the results of this analysis without outlier removal in the Figure S1).

RESULTS

When evaluating the held-out test set (n = 2,682), rating predictions generated by XLM-RoBERTa correlated positively with the human ratings (r = 0.80; CI 95% [0.79, 0.81]); see Figure 1.

DISCUSSION

Taken together, we have shown that automatically scoring responses on a scientific creativity task in German using a fine-tuned LLM is feasible and results in a very high correlation with human ratings. Leveraging the XLM-RoBERTa model, our study accurately predicts human ratings for a scientific creativity task in German. The consistency between our validation and test set performance indicates a robust training procedure and no evidence of overfitting. The achieved correlations between model predictions and human ratings are in line with other research using LLMs (e.g., Organisciak et al., 2023) and thus validate the effectiveness of these automated assessment methods. With that, our study closes a gap in research regarding automated scoring methods aiming to evaluate scientific creative thinking tasks in the German language.

Most obviously, by targeting the German language, we limited our work to this language, and we thus encourage researchers to reevaluate our approach for scientific creative thinking tasks available for other languages such as Turkish (Ayas & Sak, 2014), among others. In addition, our work is limited by the task used to assess divergent ideation in experimental tasks. Future work should extend this effort to other items aiming at this specific scientific thinking ability as well as other scientific creative thinking abilities such as divergent ideation in science tasks (Aschauer et al., 2022). Moreover, future work could investigate not only

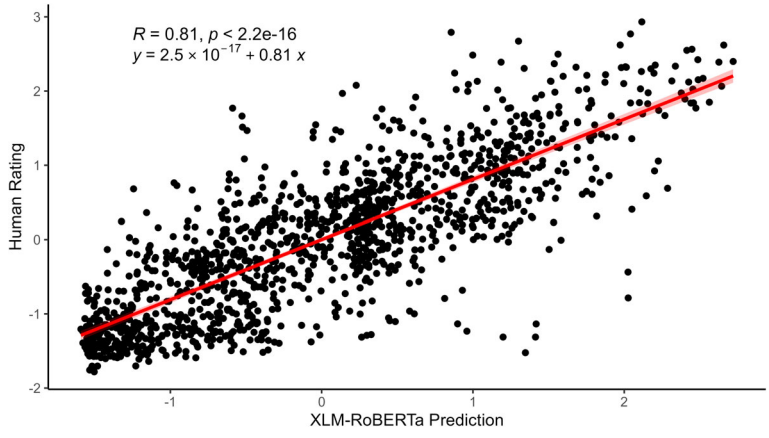

FIGURE 1. Correlation between Human-Rated Solutions and Model Predictions.*Note*. Model predictions and human ratings are *z*-transformed. $n = 2682$; 3 outliers were removed.

rater agreement but also rater disagreement (e.g., Forthmann et al., 2017), because previous work showed that raters' residual disagreement is predictive of controversial responses, which, in turn, might be considered important for broader applications, such as educational settings (Dumas et al., 2023).

Most recent efforts of automated scoring of creativity tasks emphasize the use of LLMs. However, we want to point out that competing prediction models such as xgboost (Chen & Guestrin, 2016) or random forests (Breiman, 2001), for example, have also been shown to be capable of predicting human ratings for prototypical divergent thinking tasks like the alternate uses task (Buczak, Huang, Forthmann, & Doebler, 2023). Although all of these models perform well, it is important to note their inherent differences. Unlike LLMs, prediction models like xboost or random forests do not work with plain text (i.e., natural language), but rather with meta-information (i.e., features) extracted from texts. For example, Buczak et al. (2023) constructed a set of simple text mining features such as the number of words in a response and the average word length in a response. These features are then used to inform the prediction models. These differences between models beg the question whether we could enhance our understanding of automated scoring of domain-specific creative thinking tasks in German and possibly other languages by comparing the prediction models. Future work should thus consider comparing a variety of automated scoring procedures.

Finally, we would like to point out potential issues of fairness in machine learning, for example with regards to demographic biases (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2022) or biases that might have been carried over from the training data. Although no demographic data were used to train the current LLM and the existence of reliable human biases in our data is unlikely due to the large number of raters, it cannot be fully ruled out that rating distortions exist in our data.

In conclusion, our study marks a meaningful cumulative step towards automating the assessment of domain-specific creative thinking tasks in languages beyond English, presenting validated approaches that closely mimic human raters' judgments.

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

All files and data for analyses are available at Open Science Framework: https://osf.io/aw95p.

REFERENCES

Acar, S., Organisciak, P., & Dumas, D. (2023). *A comparison of supervised and unsupervised learning methods in automated scoring of figural tests of creativity*. ResearchGate.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework*. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Pp. 2623–2631. Available from: 10.1145/3292500.3330701.

Aschauer, W., Haim, K., & Weber, C. (2022). A contribution to scientific creativity: a validation study measuring divergent problem solving ability. *Creativity Research Journal*, *34*(2), 195–212; doi: 10.1080/10400419.2021.1968656.

Ayas, M.B., & Sak, U. (2014). Objective measure of scientific creativity: Psychometric validity of the Creative Scientific Ability Test. *Thinking Skills and Creativity*, *13*, 195–205; doi: 10.1016/j.tsc.2014.06.001.

Beaty, R.E., & Johnson, D.R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), Article 2; doi: 10.3758/s13428-020-01453-w.

Bossomaier, T., Harré, M., Knittel, A., & Snyder, A. (2009). A semantic network approach to the creativity quotient (CQ). *Creativity Research Journal*, *21*(1), 64–71; doi: 10.1080/10400410802633517.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2023). The machines take over: A comparison of various supervised learning approaches for automated scoring of divergent thinking tasks. *The Journal of Creative Behavior*, *57*(1), 17–36; doi: 10.1002/jocb.559.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pp. 785–794. Available from: https://doi.org/10.1145/2939672.2939785.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale (arXiv:1911.02116). *arXiv*. Available from: http://arxiv.org/abs/1911.02116.

Cropley, D.H., & Marrone, R.L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication; doi: 10.1037/aca0000510.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). *arXiv*. Available from: http://arxiv.org/abs/1810.04805.

Dumas, D., Acar, S., Berthiaume, K., Organisciak, P., Eby, D., Grajzel, K., … Carrera, M. (2023). What makes children's responses to creativity assessments difficult to judge reliably? *The Journal of Creative Behavior*, *57*(3), 419–438; doi: 10.1002/jocb.588.

Dumas, D., & Dunbar, K.N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67; doi: 10.1016/j.tsc.2014.09.003.

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(4), 645–663; doi: 10.1037/aca0000319.

Ferrando, P.J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*(5), 762–780; doi: 10.1177/0013164417719308.

Forster, E.A., & Dunbar, K.N. (2009). *Creativity evaluation through latent semantic analysis*. Proceedings of the Annual Meeting of the Cognitive Science Society, 31. Available from: https://escholarship.org/uc/item/4wp633ph.

Forthmann, B., & Doebler, P. (2022). Fifty years later and still working: rediscovering Paulus et al.'s (1970) automated scoring of divergent thinking tests. *PsyArXiv*. Available from: https://doi.org/10.31234/osf.io/byj8c.

Forthmann, B., Goecke, B., & Beaty, R.E. (2023). Planning missing data designs for human ratings in creativity research: A practical guide. *Creativity Research Journal*, *1–12*; doi: 10.1080/10400419.2023.2250976.

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139; doi: 10.1016/j.tsc.2016.12.005.

Green, A.E., Kraemer, D.J.M., Fugelsang, J.A., Gray, J.R., & Dunbar, K.N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264–272; doi: 10.1037/a0025764.

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

Johnson, D.R., Kaufman, J.C., Baker, B.S., Patterson, J.D., Barbot, B., Green, A.E., … Beaty, R.E. (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, *55*(7), 3726–3759; doi: 10.3758/s13428-022-01986-2.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach (arXiv:1907.11692). *arXiv*. Available from: http://arxiv.org/abs/1907.11692.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35; doi: 10.1145/3457607.

Merseal, H.M., Beaty, R.E., Kenett, Y.N., Lloyd-Cox, J., De Manzano, Ö., & Norgaard, M. (2023). Representing melodic relationships using network science. *Cognition*, *233*, 105362; doi: 10.1016/j.cognition.2022.105362.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*, 1–33; doi: 10.1002/j.2333-8504.1992.tb01436.x.

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, 101356; doi: 10.1016/j.tsc.2023.101356.

Patterson, J.D., Barbot, B., Lloyd-Cox, J., & Beaty, R.E. (2023). AuDrA: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*; doi: 10.3758/s13428-023-02258-3.

Patterson, J.D., Merseal, H.M., Johnson, D.R., Agnoli, S., Baas, M., Baker, B.S., ... Beaty, R.E. (2023). Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, *17*(4), 495–507; doi: 10.1037/aca0000618.

Paulus, D.H. (1970). *Computer simulation of human ratings of creativity*. Final Report.

R Core Team. (2022). *R: A language and environment for statistical computing [Software]*. Vienna: R Foundation for Statistical Computing. Available from: https://www.R-project.org/.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, *1968*(1), 1–169; doi: 10.1002/j.2333-8504.1968.tb00153.x.

Weiss, S., Olderbak, S., & Wilhelm, O. (2023). Conceptualizing and measuring ability emotional creativity. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication; doi: 10.1037/aca0000585.

Weiss, S., Wilhelm, O., & Kyllonen, P. (2021). An improved taxonomy of creativity measures based on salient task attributes. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication; doi: 10.1037/aca0000434.

Yu, Y., Beaty, R.E., Forthmann, B., Beeman, M., Cruz, J.H., & Johnson, D. (2023). A MAD method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication; doi: 10.1037/aca0000573.

Zhou, Z.-H. (2021). *Machine Learning*. Singapore: Springer; doi: 10.1007/978-981-15-1967-3.

Zielińska, A., Organisciak, P., Dumas, D., & Karwowski, M. (2023). Lost in translation? Not for Large Language Models: Automated divergent thinking scoring performance translates to non-English contexts. *Thinking Skills and Creativity*, *50*, 101414; doi: 10.1016/j.tsc.2023.101414.

Benjamin Goecke, University of Tübingen

Paul V. DiStefano, Pennsylvania State University

Wolfgang Aschauer, Kurt Haim, University of Education Upper Austria

Roger Beaty, Pennsylvania State University

Boris Forthmann, University of Münster

Correspondence concerning this article should be addressed to Benjamin Goecke, Hector Research Institute, University of Tübingen, Tübingen, Germany. E-mail: academ@benjamin-goecke.de

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

AUTHOR NOTE

The study was conducted in accordance with the declaration of Helsinki.
The authors have no conflicts of interest to disclose.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Correlation between Human-Rated Solutions and Model Predictions without outlier removal.
Figure S2. Correlation between Human-Rated Solutions, Fluency, and Flexibility.
Table S1. Simulation Designs for the Planned Missingness Rater Design.