JOURNAL OF
Creative Behavior
Publication of the Creative Education Foundation™

PHILIP BUCZAK
HE HUANG
BORIS FORTHMANN
PHILIPP DOEBLER

The Machines Take Over: A Comparison of Various Supervised Learning Approaches for Automated Scoring of Divergent Thinking Tasks

ABSTRACT

Traditionally, researchers employ human raters for scoring responses to creative thinking tasks. Apart from the associated costs this approach entails two potential risks. First, human raters can be subjective in their scoring behavior (inter-rater-variance). Second, individual raters are prone to inconsistent scoring patterns (intra-rater-variance). In light of these issues, we present an approach for automated scoring of Divergent Thinking (DT) Tasks. We implemented a pipeline aiming to generate accurate rating predictions for DT responses using text mining and machine learning methods. Based on two existing data sets from two different laboratories, we constructed several prediction models incorporating features representing meta information of the response or features engineered from the response's word embeddings that were obtained using pre-trained GloVe and Word2Vec word vector spaces. Out of these features, word embeddings and features derived from them proved to be particularly effective. Overall, longer responses tended to achieve higher ratings as well as responses that were semantically distant from the stimulus object. In our comparison of three state-of-the-art machine learning algorithms, Random Forest and XGBoost tended to slightly outperform the Support Vector Regression.

Keywords: divergent thinking, creative quality, human ratings, supervised learning, Random Forest, gradient boosting, Support Vector Regression.

Scoring of creative thinking tasks is a laborious endeavor that requires human and time resources. Motivated by a potential reduction of scoring efforts automated scoring of creative thinking tests has become a current hot topic in creativity research. This is documented by several published recent attempts for automated scoring of the popular Alternate Uses Task (AUT; e.g., Beaty & Johnson, 2021; Dumas, Organisciak, & Doherty, 2020; Stevenson et al., 2020). Prior work has mostly relied on semantic distance approaches as a way to automatically quantify the originality of the responses. These approaches are examples of unsupervised machine learning approaches and they have shown remarkable success as evidenced by strong correlations at the person-level between semantic distance scores and scores provided by human raters (Beaty & Johnson, 2021; Dumas et al., 2020). However, (semi-)supervised learning approaches have also been successfully implemented (Stevenson et al., 2020). While all of these approaches displayed promising results, there is still room for further improvement, and the current work aims at extending and complementing existing work on automated scoring of the AUT. Specifically, this work focuses on supervised learning approaches which have been less extensively studied in this context, as well as on improved pre-processing of text data. The goal was to examine a supervised learning pipeline in terms of its rating performance using varying supervised learning algorithms (i.e., Random Forest, XGBoost, and Support Vector Regression [SVR]) as well as varying feature set compositions differing in size and interpretability. We further aimed at finding the most relevant features for the prediction of human ratings.

DIVERGENT THINKING AND ITS ASSESSMENT

Divergent thinking (DT) refers to the cognitive capacity to generate multiple options in response to a given task (Guilford, 1967). Hence, the response format is open-ended which naturally implies a scoring of

The Journal of Creative Behavior, Vol. 57, Iss. 1, pp. 17–36 © 2022 The Authors. *The Journal of Creative Behavior* published by Wiley Periodicals LLC on behalf of Creative Education Foundation (CEF). DOI: 10.1002/jocb.559

productivity (i.e., the number of responses). However, this productivity scoring (i.e., fluency) has often been criticized (e.g., Zeng, Proctor, & Salvendy, 2011) for its lack of conceptual relevance with respect to creativity. Creativity commonly refers to something that is perceived as original and useful (Runco & Jaeger, 2012) in the given context. Hence, DT responses should be scored for originality and usefulness beyond mere productivity indices such as fluency. Original responses in DT tasks can be identified based on three classical facets (Wilson, Guilford, & Christensen, 1953): uncommonness, cleverness, and remoteness. Rater instructions often contain explicit instructions to consider all three when scoring originality (e.g., Hofelich-Mohr, Sell, & Lindsay, 2016; Silvia et al., 2008).

Uncommonness refers to the statistical rarity of a response Cropley, 1967; Wallach & Kogan, 1965; Wilson et al., 1953) and is historically perhaps the oldest originality indicator (Hargreaves, 1927). Cleverness refers to the cunning aptness of a response and clever responses are maturely thought through. Cleverness has traditionally been scored by human judges (Forthmann et al., 2017; French et al., 1963; Wilson et al.,1953). Finally, remoteness refers to responses that are associatively more distant as compared to the most obvious responses (Silvia et al., 2008; Wilson et al., 1954). Traditionally, remoteness was scored for the Consequences Task (e.g., generating possible consequences for the scenario that people do not need to sleep anymore) with more far-reaching consequences being more remote than immediate ones (e.g., that people go to work at night).

In addition, remoteness has been conceptualized as semantic distance of responses in comparison to the task stimulus or a semantic representation of it (Dumas & Dunbar, 2014; Hass, 2017). This way originality can be automatically scored by semantic vector models of meaning such as Latent Semantic Analysis (Landauer & Dumais, 1997), GloVe (Pennington, Socher, & Manning, 2014), and Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013). For example, in the AUT, the semantic distance between the item object (e.g., knife) and a response (e.g., use it as a mirror) is calculated to reflect the remoteness of the response. Early validity evidence for scoring based on semantic vector models was quite inconsistent across studies (Forster & Dunbar, 2009; Harbison & Haarman, 2014; Hass, 2017). Elaboration bias (semantic distance depends technically on the number of words; Forthmann et al., 2019) of the most often used LSA approach was identified as one potential source to explain these inconsistent findings.

Newer studies, however, have successfully addressed the issue of elaboration bias and demonstrated strong validity evidence for person-level aggregated scores (e.g., Beaty & Johnson, 2021; Dumas et al., 2020). For example, Beaty and Johnson (2021) improved automatic scoring by relying on a variety of semantic spaces (created by different approaches) as well as multiplicative compositional models instead of additive ones, whereas Dumas et al. (2020) improved automated scoring by a weighting approach that relies on inverse document frequency. Notably, beyond the AUT, semantic distance scoring displayed validity evidence for the Torrance Test of Creative Thinking (Acar et al., 2021), creative verb association (Beaty & Johnson, 2021; Heinen & Johnson, 2018; Prabhakaran, Green, & Gray, 2014), the Consequences Test (LaVoie, Parker, Legree, Ardison, & Kilcullen, 2020), the Remote Associates Test (Beisemann, Forthmann, Bürkner, & Holling, 2020), or abstract figure naming (Sung, Cheng, Tseng, Chang, & Lin, 2022).

SUPERVISED MACHINE LEARNING

Supervised machine learning refers to the construction of predictive models rather than descriptive ones (Lantz, 2013). In such models, an algorithm learns to predict values for a target variable of interest as opposed to unsupervised machine learning models in which the goal is to build a descriptive model of the data, typically with the aim to abstract from the original data and discover structure. For example, the above-mentioned vector models of semantic meaning (e.g., LSA, HAL, GloVe, and Word2Vec) are examples of unsupervised learning algorithms because vector spaces are created towards a quantitative description of how different terms relate semantically to each other and not with the goal to predict any specific target. Supervised models, however, try to approximate an unknown functional relationship between covariates (also called features) and an outcome variable by learning a model from examples. Such functional relationships can originate either from a regression context where the outcome is continuous or from a classification context where the outcome is discrete and input features are mapped to a class label. Here, we consider the prediction of (mean) creative quality ratings as a regression problem.

There is a wide variety of machine learning algorithms differing in their mathematical and conceptual approaches, interpretability, numerical demands, and context of the application. In this work, we focus on Random Forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016), because they have shown to be highly performant in a variety of situations with structured tabular data, and on SVR, the regression analog of the highly competitive and somewhat more well-known Support Vector Machine (Vapnik, 2000) for classification. All three, Random Forest, XGBoost, and SVR, are known to rely on custom feature engineering for peak performance, so we will combine them with different feature sets.

USING SUPERVISED MACHINE LEARNING FOR AUTOMATED SCORING OF DT TASKS

Most of the recent attempts to automatically score DT tasks for originality rely on unsupervised learning. While the current work aims at extending the existing work on automatic scoring of DT tasks it should be noted that very few supervised learning approaches exist in the literature. First, there is the pioneering work by Paulus, Renzulli, and Archambault Jr. (1970) that has been neglected until nowadays. They used a sample of $N = 100$ participants to develop prediction equations for fluency, flexibility, and originality of the responses based on features of the response such as number of question marks, number of commas, number of periods, number of words, number of sentences, average word length, the standard deviation of word length, for example (for complete feature lists for all activities see Paulus et al., 1970). Prediction equations were developed based on stepwise multiple regression analysis and cross-validated on a sample of $N = 53$ participants. Their cross-validation results were quite impressive with correlations between automatically derived scores and human rater scores in the range from .42 to .96, and most of the correlations reported exceeded .70. More than 50 years later these findings are still impressive and promising as the computational power and availability of powerful algorithms have clearly increased over the last five decades. A more recent approach for a (semi-)supervised learning algorithm to score DT tasks has been developed by Stevenson et al. (2020). Based on the word embeddings (WE) derived from Word2Vec models, they cluster AUT task responses w.r.t. their semantic distance first. For each cluster, a representative mean creativity score is obtained by averaging all cluster-specific ratings. New responses are then assigned the creativity score of the semantically nearest cluster. Thus, Stevenson et al. (2020) employ a hybrid, semi-supervised approach that combines unsupervised learning (through clustering) and supervised learning (deriving predictions from observed examples). Their reported validity evidence was clearly on par with the unsupervised approaches proposed by Beaty and Johnson (2021) as well as Dumas et al. (2020).

AIM OF THE CURRENT STUDY

The aim of our study is to develop and evaluate an ML-based approach for automated scoring of DT tasks. In contrast to Stevenson et al.'s (2020) hybrid approach, our approach stems more from the traditional perspective in ML. More precisely, we engineer custom features from original DT data sets and employ frequently used machine learning algorithms, namely Random Forest, XGBoost, and SVR, to predict mean creativity ratings. We aim to compare the predictive performance of these three algorithms, and study the impact of different feature sets (including the influence of individual features) and semantic spaces to derive conclusions and recommendations for practical use cases.

To this end, we follow the general structure of the pipeline depicted on the right-hand side of Figure 1. Whereas in the classical human-based approach (left-hand side), human raters base their ratings directly on the raw response data, our ML-based approach requires further data processing. This usually includes some form of pre-processing, for example, removing spacing, punctuation, or encoding errors. Once the responses are properly pre-processed, they are used to generate a set of informative features. The resulting data set is then used to train the machine learning model. This step can also encompass some sort of model selection, for example, comparing different learners, performing feature selection, or hyperparameter optimization. The final model can then be used to obtain predictions for new data points.

METHODS

DATA AND FEATURE ENGINEERING

For our analysis, we used two DT data sets by Silvia et al. (2008) and Hofelich-Mohr et al. (2016), both containing rated responses from AUTs. While the former data set contains 3,432 responses to alternative uses of *brick* and *knife* rated by three human judges, the latter contains 3,870 responses to alternative uses of *brick* and *paperclip* rated by four human judges. In both cases, ratings ranged from 1 (worst) to 5 (best). For Silvia et al. (2008), inter-rater reliability was .48 (based on an absolute agreement intra-class correlation coefficient for average ratings), and for Hofelich-Mohr et al. (2016) it was .66. According to Cicchetti's (2001) criteria, the obtained levels of interrater reliability were *fair* and *good*, respectively.

Data formatting and sanitization (e.g., removing errors in spacing, punctuation, encoding, etc.) was not necessary for both data sets, because they were already pre-processed in the original source. However, text

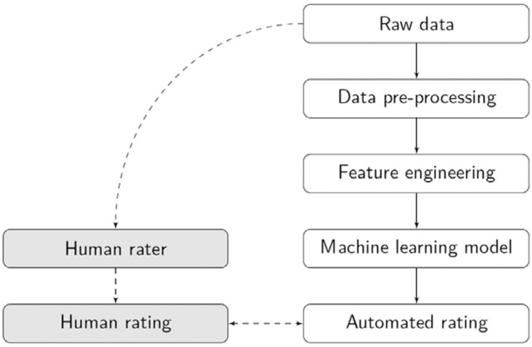

FIGURE 1. Pipeline for automated scoring of DT tasks using machine learning.

formatting and sanitization is generally an important pre-processing step. For the purpose of our analysis, we averaged the respective ratings for each observation. Thus, a baseline observation contained the participant's response, the respective stimulus item, and the response's mean rating. We generated further features which are displayed in Figure 2. The different features can be grouped based on their type and interpretability.

The feature group "meta-features" consists of the features "number of words," "average word length," and "maximum word length". These can be obtained from the respective response and are handily interpretable. The WE describe the word vector representation of a response w.r.t. to a (pre-trained) semantic space. Based on the dimensionality d of the semantic space, d features are added, each representing a word vector space component "loading". Word embeddings are not interpretable by themselves. Generally, pre-

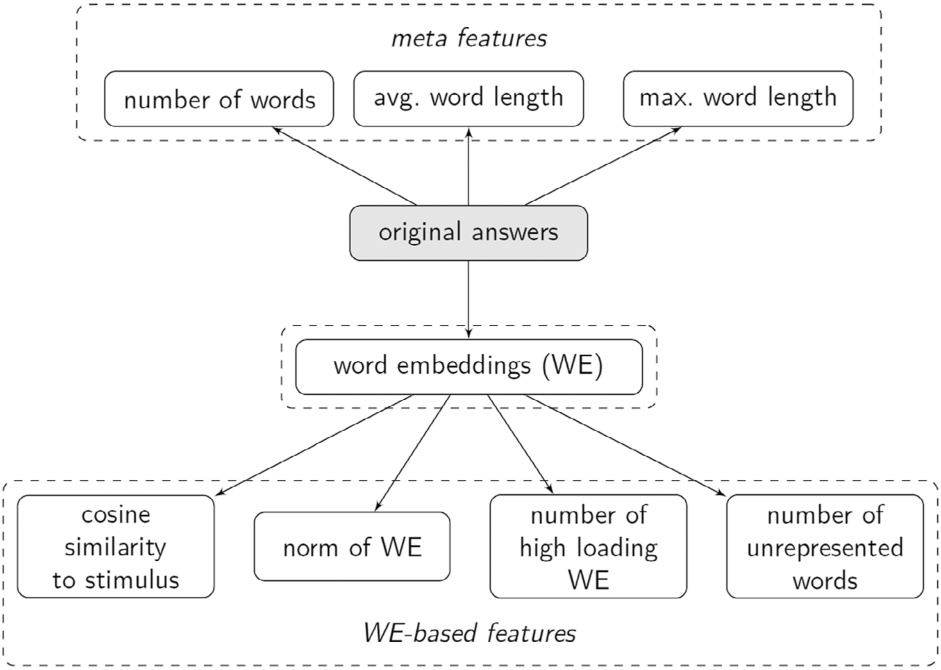

FIGURE 2. Features engineered from original answers.

trained semantic spaces provide vector representations for single words. However, responses in DT tasks are usually phrases or sentences. To obtain sentence embeddings, we sum the WE of the individual words contained in the sentence.

The feature group "WE-based features" contains the similarity of the response to the stimulus item (based on the cosine distance of the respective WE), the norm of the response's WE vector, the number of a response's "high" loading WE components (value larger than the 75%-quantile) and the number of words from the response that could not be represented using the respective semantic space (corpus missings). These features appear interpretable but one must keep in mind that they are based on uninterpretable WE.

MACHINE LEARNING ALGORITHMS

Decision trees

A major class of supervised machine learning algorithms is decision tree models such as Breiman's classification and regression trees (CART; Breiman, Friedman, Stone, & Olshen, 1984). Decision trees split the feature space into disjoint regions in which the target function is predicted through a constant value. For regression trees, the constant value is chosen as the mean observed outcome of all observations falling into the region. Decision trees can be thought of as a tree-like flow chart in which a series of yes/no decisions (splits) leads to a prediction for the outcome. An exemplary decision tree is depicted in Figure 3. In this example, an answer containing two words with an average length of five would lead to a predicted mean creativity rating of 2.5. In the CART algorithm, the splits are chosen such that the total variance of the observations falling into the two subsequent regions is minimized. Although highly interpretable, single tree models suffer from high variability as little perturbations in the data may already lead to distinctly different tree structures. Further, single tree models are often insufficient to capture complicated functional relationships, and thus, several tree models are often combined to form so-called ensembles.

Random Forest

Random Forests (RF; Breiman, 2001) are an ensemble method comprised of a large number of decision trees whose individual predictions are combined into an overall prediction. The RF algorithm grows its trees independently on bootstrap samples of the original data (Efron & Tibshirani, 1993). The RF prediction is then determined (in a regression context) by averaging over all individual tree predictions. It can be shown that the variance of the RF prediction depends on the pairwise correlation of the individual trees as well as the number of trees (Hastie, Tibshirani, & Friedman, 2009). The greater the number of trees and the smaller the pairwise correlation, the smaller the variance term is. Thus, the number of trees is usually chosen large (i.e., several hundred). As for the pairwise correlation, the RF algorithm aims to de-correlate its trees by restricting the number of split variables considered inside the individual tree models and instead choosing a random subset of potential split variables. In this sense, a forest of individual trees with a restricted set of variables is grown. Thus, the RF reduces the high variability of individual tree models overall, but at the same time loses the high interpretability of single trees as its prediction is based on aggregating several hundred trees.

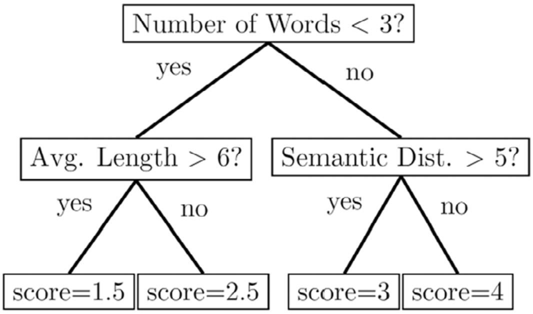

FIGURE 3. An example of a single decision tree.

Gradient Boosting Decision Trees

Gradient Boosting Decision Trees (GBDT; Friedman, 2001) are another ensemble method based on decision trees. Similar to RFs, GBDT reduce the high variability of single tree models but offers only limited interpretability. Different from RFs, however, GBDT models are generated in a stepwise fashion aimed at reducing the mismatch between prediction and data by iteratively fitting decision trees to the discrepancy between prediction and data. A large number of trees can make the discrepancy arbitrarily small but increase the risk of overfitting, that is, the model would resemble the training data too closely and fail to successfully predict unknown future data (Hastie et al., 2009). In this work, we use XGBoost (XGB; Chen & Guestrin, 2016) which is a highly performant and fast implementation of GBDT adding various regularization techniques to reduce the risk of overfitting.

Support Vector Regression

In contrast to regular least squares regression which optimizes a squared loss, the SVR (Vapnik, 2000) uses the so-called ϵ-insensitive loss (see Appendix A) which only penalizes errors of magnitude greater than a pre-specified ϵ and ignores smaller errors. The difference between these two loss functions is visualized in Figure A1 (see Appendix A). Geometrically speaking, the SVR aims to place a tube in the data space that encloses as many data points as possible (i.e., minimizing the ϵ-insensitive loss) while at the same time penalizing too wide tube diameters. SVR can also be used to model non-linear functions by transforming the observations into higher dimensional spaces using non-linear mappings and finding an optimal solution there. To avoid costly computations in high-dimensional spaces, so-called kernel functions are used to perform the required calculations in the original (lower-dimensional) space rather than the transformed space which can greatly reduce the computational complexity. We refer to Vapnik (2000) for examples of such kernel functions.

DATA ANALYSIS

We implemented our pipeline in R (R Core Team, 2020). Machine learning experiments were conducted with the mlr (Bischl et al., 2016) interface using the RF, SVR, and XGB implementations from the ranger (Wright & Ziegler, 2017), e1071 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2020), and xgboost (Chen et al., 2020) package, respectively. For our analysis, we considered four different feature sets varying in size and interpretability as well as five different semantic spaces.

Table 1 provides an overview of the feature sets. Out of these, only the second feature set does not depend on the WE. Containing only the number of words, average word length, and maximum word length, it is the simplest and most interpretable feature set, in stark contrast to the first feature set which only contains the WE and is, thus, completely uninterpretable. The third feature set adds further features which, however, are only interpretable to a certain degree since they are based on the WE. Feature Set 4 combines all features, i.e. meta information, WE-based features, and the WE.

The pre-trained semantic spaces we used for our analysis differed in word count and dimensionality. The word count refers to the number of words for which the semantic space contains vector representations. The dimensionality of the semantic space refers to the length of the word vector representations. The semantic spaces we used, are further specified in Table 2. Four were originally obtained using the GloVe algorithm (Pennington et al., 2014), while the fifth was obtained through the Word2Vec method (Mikolov et al., 2013). For determining the WE of the answers, we considered additive composition and zero padding in our analysis.

TABLE 1. Feature Sets Used for Analysis. Cf. Figure 2 for the Explanation of Feature Groups

Feature set	Included feature groups
1	WE
2	Meta information
3	Meta information + WE-based features
4	Meta information + WE-based features + WE

Note. WE, Word embeddings.

TABLE 2. Semantic Spaces Used for Analysis

Algorithm	Words	Dimensionality
GloVe	400,000	50
GloVe	400,000	100
GloVe	400,000	300
GloVe	2,000,000	300
Word2Vec	3,000,000	300

Our analysis approach was 2-fold. Comparison Study 1 compared the predictive performance of RF, SVR, and XGB models for the presented data sets. The aim of Comparison Study 2 was to analyze the generalizability of the ML models using "external" data for validation. For this purpose, we used one of the data sets for training (i.e., fitting) the models and the other data set, respectively, for validation. Since the data sets for training and validation originated from different data generating processes, the performance results gave an indication of how well these models could generalize.

All three ML algorithms used in our analysis have an individual set of hyperparameters that require tuning. Thus, we combined our model training with hyperparameter tuning. Table A1 (see Appendix A) contains an overview of the hyperparameter sets and respective search spaces used for tuning. To achieve an honest comparison between our algorithms in Comparison Study 1, we employed a nested resampling approach using 5-fold cross-validation (CV) in the outer validation and a 3-fold CV in the inner hyperparameter tuning loop. This is needed because tuning and validating the same data instances would lead to overly optimistic error estimates (Cawley & Talbot, 2010). For Comparison Study 2, we performed the hyperparameter tuning via a 10-fold CV on the training data and used the optimal parameter choices for fitting the respective models on the entire training data set.

RESULTS

COMPARISON STUDY 1

Figure 4 shows the RMSE values achieved in 100 replications of nested resampling for a given combination of data set, learner, and feature set with an additive composition of WE. Because the choice of semantic space did not have noticeable impact on the performance, we are only showing the results for the semantic space GloVe 6B 50d. We refer to the Online Supplement for the results for the remaining semantic spaces. As a benchmark, we have provided the mean RMSE obtained when using the mean rating computed from the training data set as a constant prediction for all responses in the validation data set.

For both data sets, the predictive performance was worst when relying only on meta information from the answer. Including the information contained in the WE either indirectly (through WE-based features) or directly, greatly improved the performance. The best RMSE values were generally reached by using all features combined. Comparing the three ML learners, the RF models seemed to perform best in most of the scenarios tying with XGB in some cases. There is no scenario in which the SVR achieved the lowest mean RMSE.

These results are echoed when looking at the correlation between automated predictions and the respective human ratings on a response level as shown in Figure 5. The highest response-level correlations scores were achieved with the combination of all features using RF ranging from .71 to .73 on the Hofelich-Mohr et al. (2016) data set and from .55 to .57 on the Silvia et al. (2008) data set.

We obtained similar results when computing correlations on the person-level. Figure 6 shows that RF using the combined feature set performs best, achieving scores from .72 to .75 for the Hofelich-Mohr et al. (2016) data and from .58 to .65 for the Silvia et al. (2008) data.

COMPARISON STUDY 2

Figure 7 shows the RMSE values achieved in the cross-sample analysis. As in Comparison Study 1, the choice of semantic space did not have a noticeable impact on the performance results. When using the Hofelich-Mohr et al. (2016) data for model training and the Silvia et al. (2008) data for validation (A), the ML models could not outperform the mean RMSE achieved when using the mean rating from the training data as the prediction for the validation data. Thus, the ML models did not generalize well in this case. Only the SVR could improve upon the benchmark when using the WE as feature. In contrast, when using the

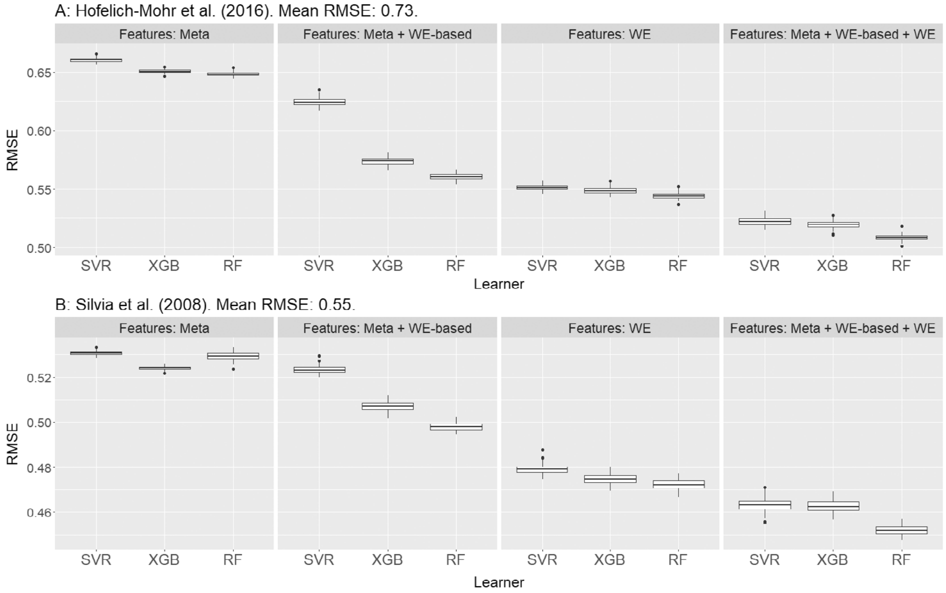

FIGURE 4. Nested resampling RMSE values for Hofelich-Mohr et al. (2016) data (a) and Silvia et al. (2008) data (b).

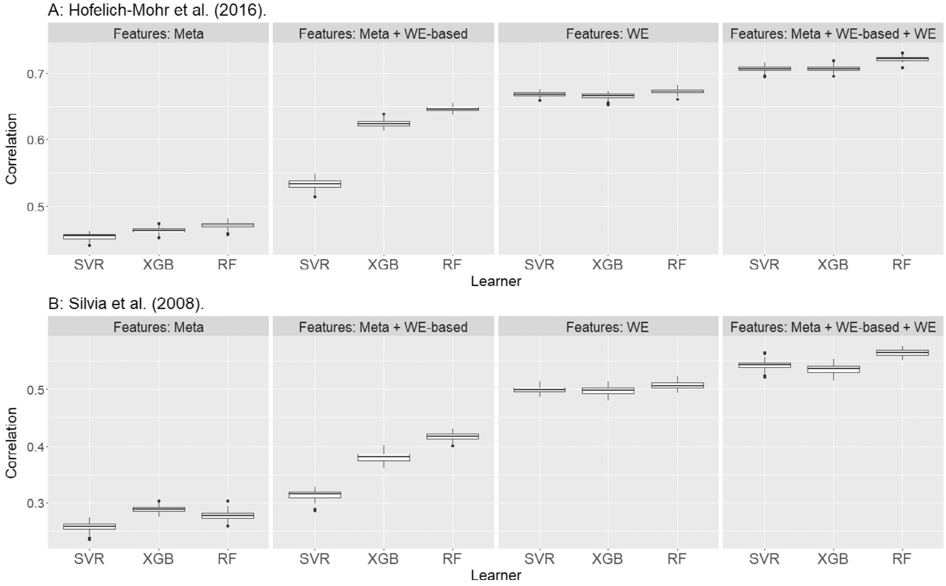

FIGURE 5. Nested resampling response-level correlations for Hofelich-Mohr et al. (2016) data (a) and Silvia et al. (2008) data (b).

Hofelich-Mohr et al. (2016) data for model training and the Silvia et al. (2008) data for validation (B), the ML models outperformed the benchmark mean RMSE in all scenarios indicating improved generalizability. The lowest RMSE values were reached by XGBoost using the combined feature set.

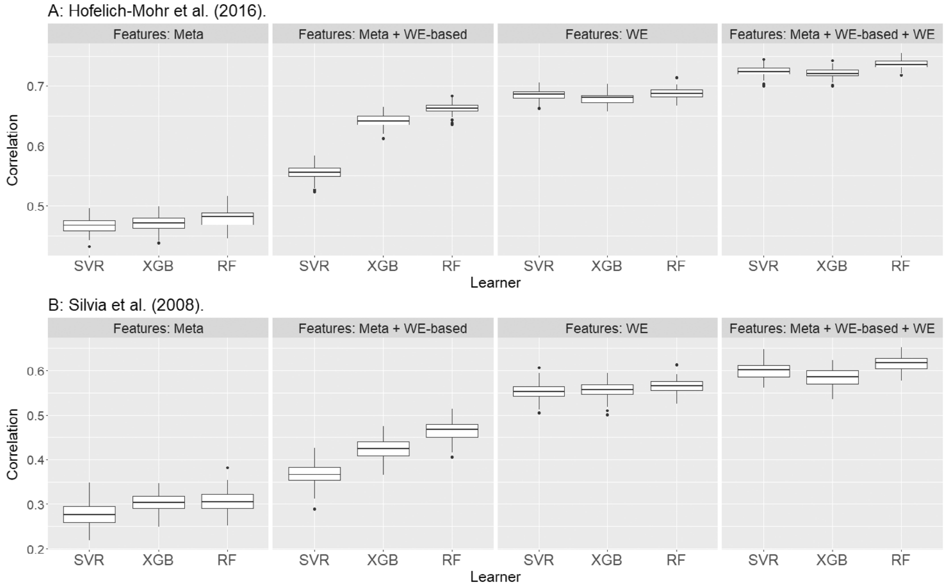

FIGURE 6. Nested resampling person-level correlations for Hofelich-Mohr et al. (2016) data (a) and Silvia et al. (2008) data (b).

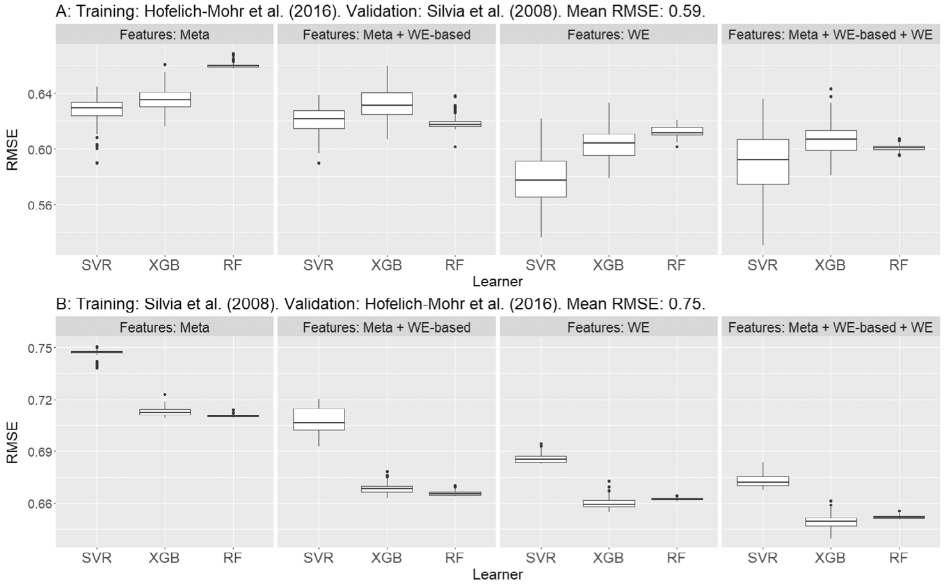

FIGURE 7. Cross-sample RMSE values.

Regarding the response-level correlations (Figure 8), the three ML performed similarly well once WE were included as features. The highest response-level correlation scores were achieved with the combined feature set. When validating the Silvia et al. (2008) data (A) RF reached correlation scores between .57 and

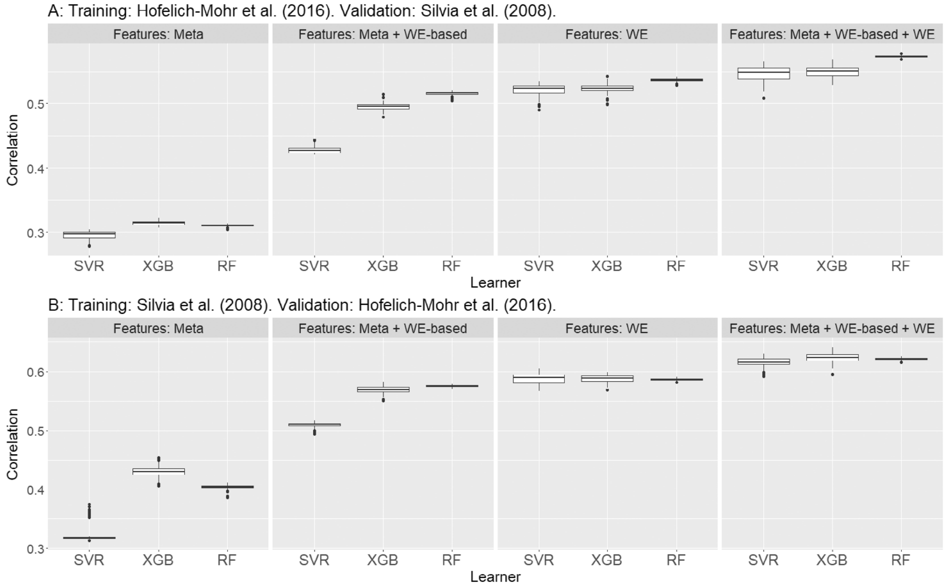

FIGURE 8. Cross-sample response-level correlations between predicted and observed mean ratings.

.58. Interestingly, despite performing best w.r.t. to the RMSE, the SVR achieved lower mean correlations than the RF and XGB models. When using the Hofelich-Mohr et al. (2016) data for validation (B), XGBoost achieved the highest correlation scores ranging between .59 and .64.

On the person-level (Figure 9), RF and SVR performed similarly when validating the Silvia et al. (2008) data (A) using the combined feature set. RF reached correlation scores of .76 to .78. When validating the data from Hofelich-Mohr et al. (2016), RF achieved correlation scores between .59 and .60.

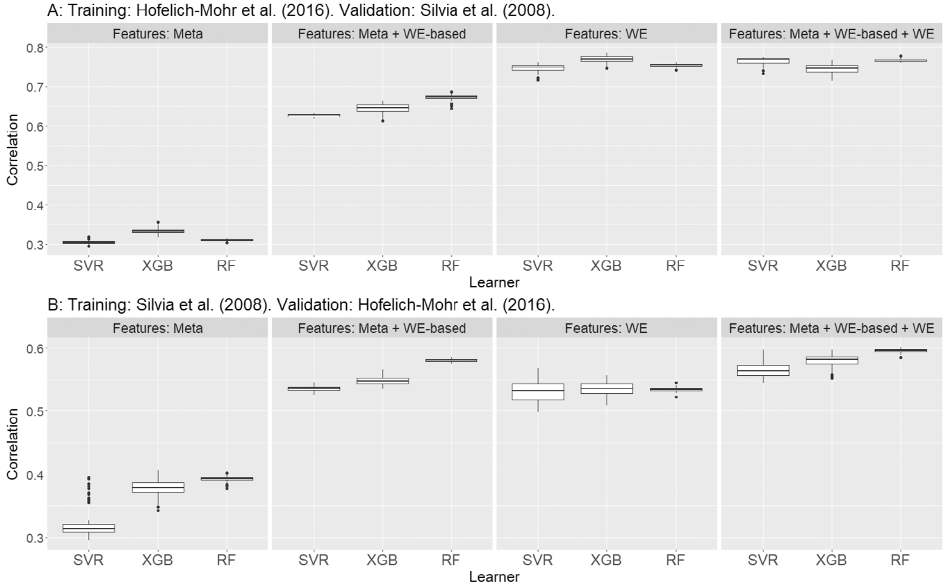

FIGURE 9. Cross-sample correlations between predicted and observed mean ratings on the person-level.

FEATURE IMPORTANCE AND PARTIAL DEPENDENCE

To analyze how predictive individual features were, we computed variable importance scores for RF and XGB models trained on the Hofelich-Mohr et al. (2016) data (we observed similar findings when using the data from Silvia et al., 2008). Figure 10 shows the feature importance of the seven meta and WE-based features as well as the seven most important features for the combined feature set.

For both, RF and XGB, the stimulus similarity was the most important feature followed by the number of words. As for the combined feature set, the first seven dimensions of the WE (d1–d7) were the most important features in RF and XGB models. This seems plausible as the WE used here are the results of a Principal Component Analysis (PCA) of the original WE (see Raunak, Gupta, & Metze, 2019). The deeper mathematical reason is that the first components of a word vector space, by construction, explain more variance in the semantic space than components with a higher index.

To explore how some of these features affected the prediction, we calculated the partial dependence which indicates how the predictions partially depend on the values of the features (see Friedman, 2001). Figure 11 shows the partial dependence of the most important two (interpretable) features, that is, the stimulus similarity and the number of words. We obtained similar results for RF and XGB. As seems intuitive, the mean rating increases with the number of words used in response and decreases the more similar the response and the stimulus object become.

DISCUSSION

We compared the predictive performance of three ML algorithms, that is, RF, XGB, and SVR for the automated prediction of mean creative quality ratings in DT tasks (Guilford, 1967). These algorithms were embedded within a pipeline which also encompassed the generation of meaningful features from the original data. The features generated ranged from interpretable meta information, for example, number of words or maximum word length, to uninterpretable features such as each response's WE into a semantic space. The semantic spaces used in this work were pre-trained using GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013). In each case, the WE added several hundred potential covariates that were either used as features themselves or as a source of generating further features such as the cosine similarity between the response and the stimulus object, the WEs' norm, or the number of high loading WE components.

Our analysis showed mostly subtle differences between the ML algorithms. In most cases, RF and XGB tied for the best performance while slightly outperforming the SVR. When working with a single data source, all three algorithms significantly outperformed the RMSE benchmark for predicting the mean rating

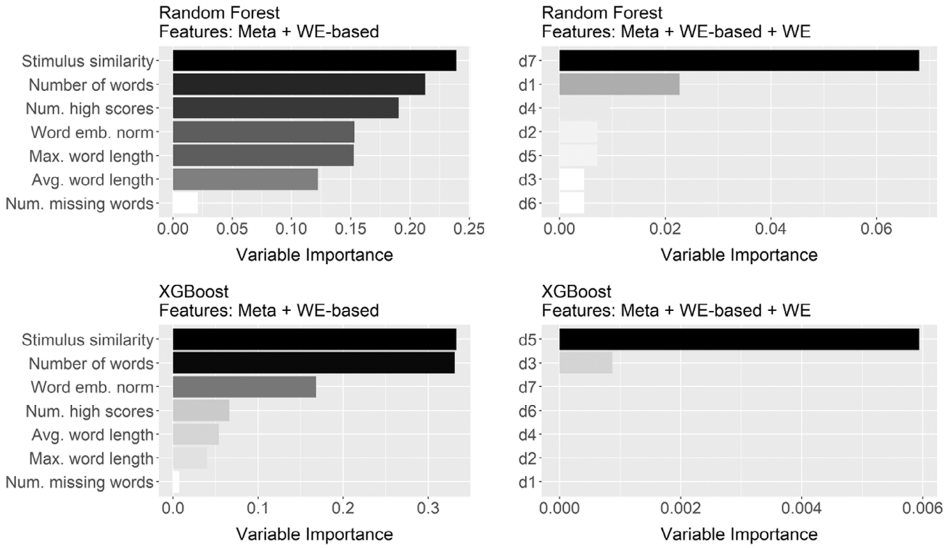

FIGURE 10. Variable importance for RF and XGB models using different feature sets.

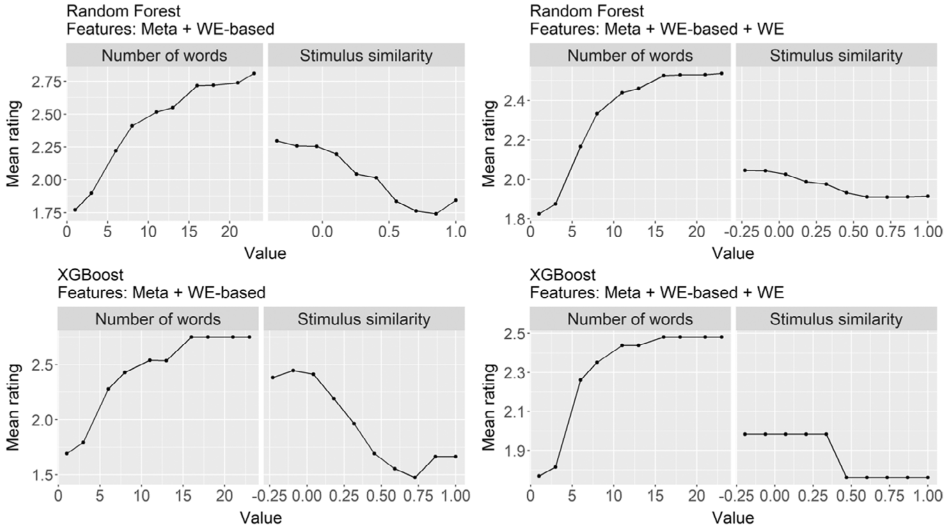

FIGURE 11. Partial dependence of stimulus similarity and number of words using RF and XGB models.

from the training set. The best model, an RF using meta information, WE as well as WE-based features, achieved person-level correlations from .72 to .75 for the Hofelich-Mohr et al. (2016) data and .58 to .65 for the Silvia et al. (2008) data. Correlations of similar size were observed cross-sample, with large linear associations of predictions from one sample to the other.

While the cross-sample correlations were satisfying, the results were not clear-cut for predicting the original rating in terms of the RMSE: When training on the Silvia et al. (2008) data and validating on the Hofelich-Mohr et al. (2016) data, all models handily outperformed the benchmark of predicting the training sample mean. Once the training/validation order is reversed, that is, when training on the Hofelich-Mohr et al. (2016) data set and validating on the Silvia et al. (2008) data set, almost all models perform worse than simply predicting the training sample mean. This indicates that the models did not generalize well to another sample in these cases, and the deeper reason behind this was mean rating differences. The qualitative differences between the RMSE and correlations are explained as follows: the means of average human ratings are hard to predict cross sample, introducing a bias from the point of view of the original ratings. Mathematically, the (square of the) bias is part of the RMSE, explaining why the simple baseline model predicting the training sample mean is competitive in terms of RMSE. In combination with the positive correlations of moderate to large size, this suggests that differences in how the 5-point Likert scales are used by raters are at the core of the cross-sample prediction challenges. Indeed, rater means for the brick stimulus vary substantially (Hofelich-Mohr et al., 2016: $M_1 = 2.29$, $M_2 = 2.58$, $M_3 = 1.64$, $M_4 = 1.61$, pooled $M = 2.03$; Silvia et al., 2008: $M_1 = 1.89$, $M_2 = 1.08$, $M_3 = 2.13$, pooled $M = 1.70$). The differing rating behaviors are further visualized in Figure A2 (see Appendix A). One way to reduce the mean shift problem is to change the prediction task: When first z-standardizing the (training) sample, the (test) sample mean is (approximately) zero. However, no faithful prediction of the original rater behavior results.

Even when assuming instruction and stimulus were completely identical in our cross-sample analysis of predictive performance, there are two inherent substantial generalizability challenges: The underlying population and the rater training might differ between studies. Both potential differences affect the rater's use of the Likert scale. While this is obvious for effects of training, population differences will for example lead to different observed maximum performance and are likely to change rater's perception of relative differences in latent ability. Relatedly, a drift of population or maturation creates a similar challenge. Any algorithmic approach, including those studied here, cannot generalize well if target population scaling differs. Mapping scales is challenging because identical benchmark responses would need to be available. Empirically, we find some overlap between responses to the brick stimulus in the two data sets, but they are mostly limited to lowly rated ideas (e.g., building a house, usage as a door stopper, or as a weapon). In sum, generalizability challenges limit the quality of predictions in other samples, but part of the problem is not specific to algorithmic approaches, as human raters trained differently would equally distort ratings in another study. In other words, an algorithmic scoring will implicitly use the scale of the original human raters.

Another potential source of bias related to the raters may be demographic bias, for example, with regard to respondents' sex or race. Demographic variables might be known to raters, and bias ratings, or merely reflected in the choice of wording of otherwise identical ideas. This also ties into the debate of fairness in ML (for a review see e.g., Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2022). Since the data used here did not contain any demographic information, we could not analyze this form of bias. Even if rater bias is not as prevalent, an automated rating algorithm may still inherit bias from the semantic space used for obtaining the WEs (Lauscher & Glavaš, 2019).

Regarding the features used, we found that including WE in some form (either directly or indirectly) greatly increased the predictive performance of the models. Models that solely contained the interpretable meta-information features were not competitive in comparison. Thus, there exists some form of trade-off between predictive performance and feature interpretability that needs to be addressed in respective applications. Additional preliminary results from our simulations also suggest that is possible to reduce the dimensionality of feature sets by performing a PCA on the WE matrix and replacing the WE components by a set of principal components. When including the first 50% of the principal components of the WE instead of all WE components themselves, our models achieved similar results as shown in Figure A3 (see Appendix A). Thus, it is possible to further reduce the computational complexity of our models without notably sacrificing predictive performance. For practical purposes, it would be interesting to study how much further the complexity of our models can be reduced before suffering a significant performance loss.

The choice of the semantic space used for generating the WEs did not have an impact on the model performance. However, further preliminary results suggest that the WE composition method may be a potential source of performance gain. In this work, we have used additive composition for determining the WE of sentences, that is, the WE of individual words were simply added up. The downside of this approach is that through the summation syntactic or word-order information of the phrases is lost (Landauer, 2002), and since not all phrases are of equal length, simply stacking the vector representation of all words together would result in different lengths of input vectors for the ML models. A simple solution is the so-called zero padding method, which adds zeros at the end of the short responses to match the length of long responses. The drawback of this method is that the dimensionality of the input vector can become particularly large for a few particularly long responses. Figure A4 (Appendix A) shows that padding improves predictive performance in most cases. One potential explanation is that additive composition entails information loss by aggregation and that several responses could lead to a similar composite. For practical use, however, it must be stressed that padding leads to greatly increased computational effort in time and memory. Therefore, if one is willing to sacrifice a small amount of predictive performance, using additive word composition may already suffice.

Overall, this work serves as proof of concept for automatically predicting creative quality ratings from DT tasks in the spirit of Paulus et al. (1970) leaving open many potential research avenues for future work. A research question of particular practical importance might be analyzing whether or not algorithms for automated predictions can distinguish between responses that would be deemed creative and responses that would be deemed nonsensical by the human rater. This could be seen as a specific instance of so-called "adversarial examples" (Biggio & Roli, 2018), an approach widely used in pattern recognition to gain insights about a model and to understand when predictions of a model tip in an unexpected direction.

The supervised approaches discussed here attempt to algorithmically reproduce human ratings, the current gold standard in AUT scoring. In contrast, unsupervised approaches rely on WEs and related features and are not explicitly optimized to reproduce human ratings, but are found to be correlated (Beaty & Johnson, 2021; Dumas et al., 2020). The semi-supervised approach of Stevenson et al. (2020) relies on distances to clusters learned in an unsupervised manner from a training set and predicts mean cluster ratings without employing regression or prediction models, circumventing the need for extensive parameter tuning. We see several clear methodological increments relative to the conceptually closest semi-supervised approach by Stevenson et al. Next to the cross-sample validation in two large English language samples, our preprocessing pipeline has the potential to improve upon all other existing approaches (unsupervised or semi-supervised). Further, our nested cross-validation framework allows us to easily change the learners (any supervised algorithm could replace the three approaches we study), the evaluation metrics, as well as the cross-sample comparison.

A logical next step is the direct comparison of unsupervised, supervised, and semi-supervised approaches. Toward a fair comparison, measures like the (cross-validated) RMSE potentially favoring supervised approaches would need to be complemented by validity evidence and robustness analysis, including adversarial examples and measures for the ability to handle previously unobserved responses.

The field of creative thinking research is currently on the mission to further improve the automated scoring of tasks such as the AUT. The current work extends and complements the existing body of research by examining supervised learning algorithms and a variety of features. The best-performing algorithms (i.e., RF and XGB) in our work push person-level correlations into the range of previous works using unsupervised or semi-supervised algorithms (Beaty & Johnson, 2021; Dumas et al., 2020; Stevenson et al., 2020), which renders them as promising algorithms to further improve automated scoring in future work (with the take-home message that the RMSE is hard to do well cross-sample because of mean differences). In addition, the studied feature sets look promising, and we have shown that dimensionality reduction approaches can help in reducing model complexity. Overall, we expect that a combination of unsupervised, semi-supervised, and supervised algorithms has the potential to push the correlation between ratings and predictions towards the correlation of two independent groups of raters, with latent variable correlation approaching unity.

CONFLICT OF INTEREST

We have no known conflict of interest to disclose.

AUTHOR CONTRIBUTIONS

The authors made the following contributions. Philip Buczak: conceptualization, formal analysis, methodology, software, visualization, original draft preparation, review, and editing; He Huang: conceptualization, formal analysis, methodology, software, visualization, original draft preparation, review, and editing; Boris Forthmann: conceptualization, funding acquisition, original draft preparation, review, and editing; and Philipp Doebler: conceptualization, funding acquisition, project administration, original draft preparation, review, and editing.

DATA AVAILABILITY STATEMENT

All R scripts used in this work are openly available in a repository of the Open Science Framework (https://osf.io/4sbhn/). Both datasets are also openly available. Data from Hofelich-Mohr et al. (2016) are available at https://conservancy.umn.edu/handle/11299/172116. Data from Silvia et al. (2008) are available at https://osf.io/8vrck/.

REFERENCES

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C., & Organisciak, P. (2021). Applying automated originality scoring to the verbal form of Torrance tests of creative thinking. *Gifted Child Quarterly*. doi: 10.1177/00169862211061874

Beaty, R.E., & Johnson, D.R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780.

Beisemann, M., Forthmann, B., Bürkner, P.-C., & Holling, H. (2020). Psychometric evaluation of an alternate scoring for the remote associates test. *The Journal of Creative Behavior*, *54*(4), 751–766.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*, 317–331.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … & Jones, Z.M. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research*, *17*(170), 1–5.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 123–140.

Breiman, L., Friedman, J., Stone, C.J., & Olshen, R.A. (1984). *Classification and regression trees*. Boca Raton, FL: CRC Press.

Cawley, G.C., & Talbot, N.L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Chen, T., & Guestrin, C. (2016). XGboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY: Association for Computing Machinery.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … & Li, Y. (2020). *xgboost: Extreme gradient boosting*. R package version 1.2.0.1. Available from: https://CRAN.R-project.org/package=xgboost

Cicchetti, D.V. (2001). Methodological commentary the precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 695–700.

Cropley, A.J. (1967). *Creativity*. London, UK: Longmans.

Dumas, D., Organisciak, P., & Doherty, M. (2020). 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.

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

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Forster, E.A., & Dunbar, K.N. (2009). Creativity evaluation through latent semantic analysis. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 602–607). Stroudsburg, PA: Cognitive Science Society.

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*, 257–269; doi: 10.1080/10400419.2017.1360059

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *Journal of Creative Behavior*, *53*(4), 559–575.

French, J.W., Ekstrom, R.B., & Price, L.A. (1963). *Manual for kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

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

Harbison, H.J.I., & Haarman, H. (2014). Automated scoring of originality using semantic representations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of COGSCI 2014* (pp. 2327–2332). Austin, TX: Cognitive Science Society.

Hargreaves, H.L. (1927). *The 'factulty' of imagination*. London, UK: Cambridge University Press.

Hass, R.W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd edn). New York, NY: Springer.

Heinen, D.J.P., & Johnson, D.R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 144–156.

Hofelich-Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, *34*(3), 347–359.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240; doi: 10.1037/0033-295X.104.2.211

Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of Learning and Motivation*, *41*, 43–84.

Lantz, B. (2013). *Machine learning with R*. Birmingham, UK: Packt Publishing.

Lauscher, A., & Glavaš, G. (2019). Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. In R.F. Mihalcea (Ed.), *Lexical and Computational Semantics (*SEM) – Proceedings of the Eighth Conference: June 6–7, 2019, Minneapolis: NAACL HLT 2019* (pp. 85–91). Stroudsburg, PA: Association for Computational Linguistics.

LaVoie, N., Parker, J., Legree, P.J., Ardison, S., & Kilcullen, R.N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399–414.

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.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2020). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071)*. R package version 1.2.0.1. Available from: https://CRAN.R-project.org/pace1071

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.

Paulus, D.H., Renzulli, J.S., & Archambault, F.X., Jr. (1970). *Computer simulation of human ratings of creativity*. *Final report*. Storrs, CT: School of Education, University of Connecticut.

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Stroudsburg, PA: Association for Computational Linguistics.

Prabhakaran, R., Green, A.E., & Gray, J.R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, *46*(3), 641–659.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raunak, V., Gupta, V., & Metze, F. (2019). Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 235–243). Stroudsburg, PA: Association for Computational Linguistics.

Runco, M.A., & Jaeger, G.J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96.

Silvia, P., Winterstein, B., Willse, J., Barona, C., 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, and the Arts*, *2*, 68–85.

Stevenson, C., Smal, I., Baas, M., Dahrendorf, M., Grasman, R., Tanis, C., … & van der Maas, H. (2020). *Automated AUT scoring using a big data variant of the consensual assessment technique: Final technical report*. Psychology Research Institute, University of Amsterdam.

Sung, Y.-T., Cheng, H.-H., Tseng, H.-C., Chang, K.-E., & Lin, S.-Y. (2022). Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000450

Vapnik, V.N. (2000). *The nature of statistical learning theory* (2nd edn). New York: Springer.

Wallach, M.A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction*. New York, NY: Holt, Rinehart & Winston.

Wilson, R.C., Guilford, J.P., Christensen, P.R. (1953). The measurement of individual differences in originality. *Psychological Bulletin*, *50*(5), 362–370. doi: 10.1037/h0060857

Wilson, R.C., Guilford, J.P., & Christensen, P.R., & Lewis, D.J. (1954). A factor-analytic study of creative-thinking abilities. *Psychometrika*, *19*(4), 297–311.

Wright, M.N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Zeng, L., Proctor, R.W., & Salvendy, G. (2011). Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity? *Creativity Research Journal*, *23*(1), 24–37.

Philip Buczak, He Huang, TU Dortmund University

Boris Forthmann, University of Münster

Philipp Doebler, TU Dortmund University

Correspondence concerning this article should be addressed to Philipp Doebler, Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany. E-mail: doebler@statistik.tu-dortmund.de

ACKNOWLEDGMENT

Open Access funding enabled and organized by Projekt DEAL.

AUTHOR NOTE

P.B. and H.H. have contributed equally to this work and are listed here in alphabetical order. The work of P.B. and H.H. is supported by a grant from The Eric & Wendy Schmidt Fund for Strategic Innovation. The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

APPENDIX A

ϵ-INSENSITIVE LOSS

For a pre-specified ϵ, the ϵ-insensitive loss $L\left(|y-\hat{f}(\mathbf{x})|_{\epsilon}\right)$ is given by

$$L\left(|y-\hat{f}(\mathbf{x})|_{\epsilon}\right) = \begin{cases} 0, & \text{if } |y-\hat{f}(\mathbf{x})| \leq \epsilon, \\ |y-\hat{f}(\mathbf{x})|-\epsilon, & \text{otherwise} \end{cases}$$

where y is the observed target value, $\mathbf{x}$ is the input feature vector and $\hat{f}(\mathbf{x}) = \hat{y}$ is the predicted target value (Vapnik, 2000).

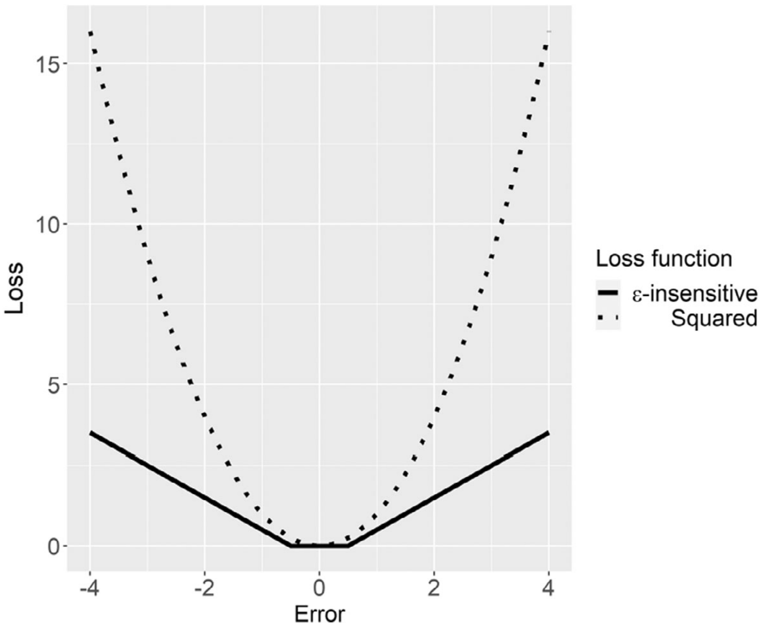

FIGURE A1. Comparison of ϵ-insensitive loss and quadratic loss.

LINEAR SVR OPTIMIZATION PROBLEM

The linear SVR optimization problem is given by finding parameters **w** and b such that

$$\Phi\left(\mathbf{w}, \boldsymbol{\xi}_\mathbf{1}, \ldots, \boldsymbol{\xi}_\mathbf{n}, \boldsymbol{\xi}_\mathbf{1}^*, \ldots, \boldsymbol{\xi}_\mathbf{n}^*\right) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right)$$

is minimized subject to

$$\begin{cases} y_i - (\mathbf{w} \cdot \mathbf{x}_i) - b & \leq \epsilon + \xi_i, \quad i = 1, \ldots, n \\ (\mathbf{w} \cdot \mathbf{x}_i) + b - y_i & \leq \epsilon + \xi_i^*, \quad i = 1, \ldots, n \\ \xi_i, \xi_i^* & \geq 0, \quad i = 1, \ldots, n \end{cases}$$

where C is a pre-specified penalization/cost factor and ξ_i, ξ_i^* are so-called slack variables for each unit i introduced to ease the optimization process, that is, to allow for observations to lie outside the ϵ-tube (Vapnik, 2000).

TABLE A1. Hyperparameters and Respective Search Spaces

Learner	Hyperparameter	Search space
Random Forest	mtry	{2, . . ., #Features}
	min.node.size	{1, . . ., 10}
	splitrule	{variance, extratrees}
Support Vector Regression	cost	2^x with $x \in [-5, 5]$
	gamma	2^x with $x \in [-5, 5]$
XGBoost	nrounds	{10, . . ., 200}
	max_depth	{1, . . ., 20}
	eta	[0.05, 0.3]
	alpha	[0, 1]
	lambda	[0, 1]
	gamma	[0, 5]

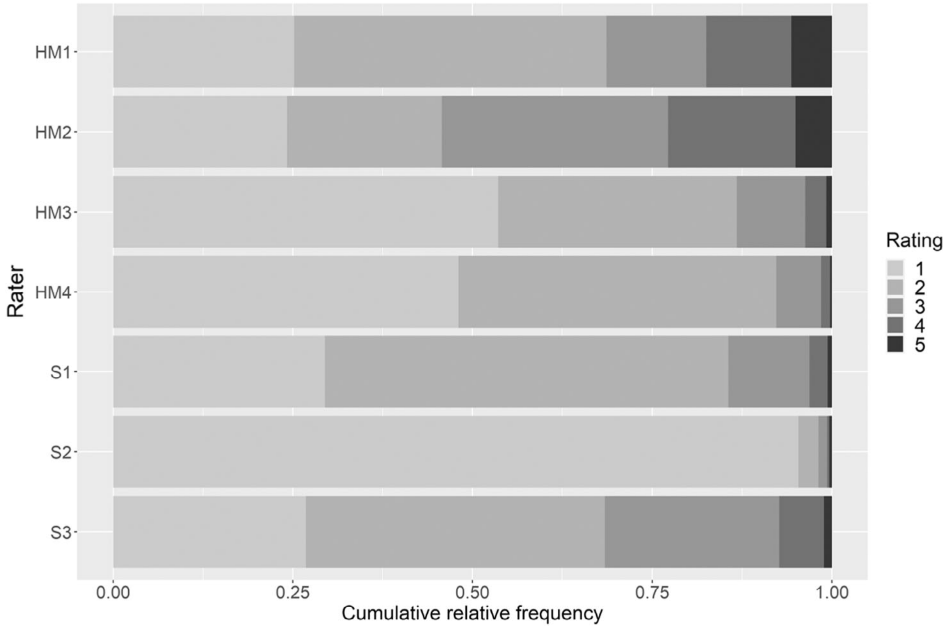

FIGURE A2. Rating behavior for the four raters (denoted by HM1–HM4) from Hofelich-Mohr et al. (2016) and the three raters (denoted by S1–S3) from Silvia et al. (2008).

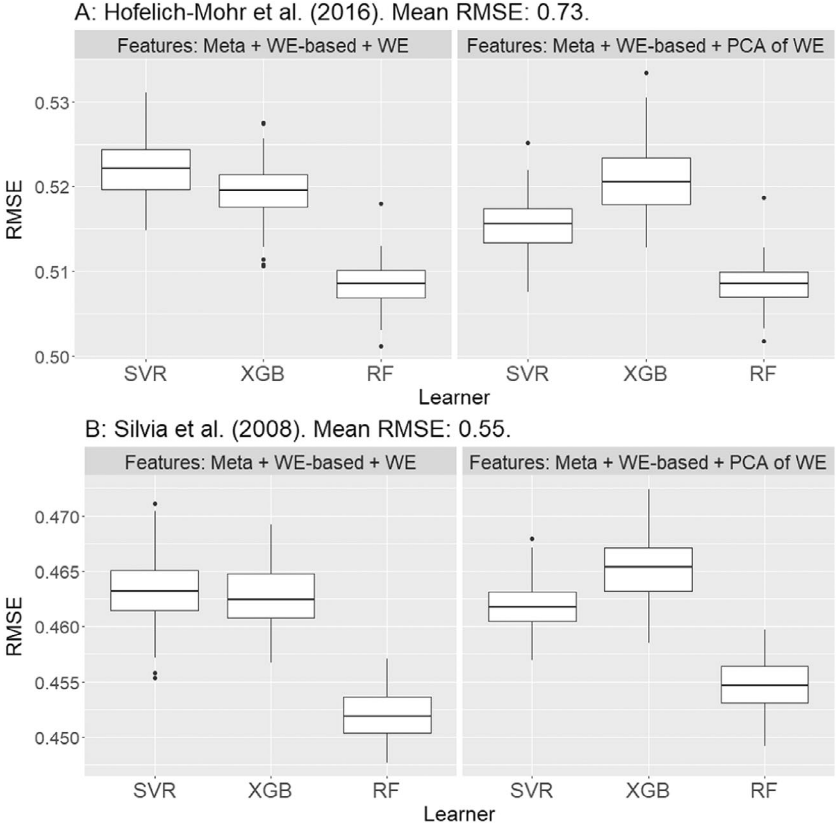

FIGURE A3. Nested resampling RMSE values when including WE or the first 50% principal components of a PCA performed on the WE.

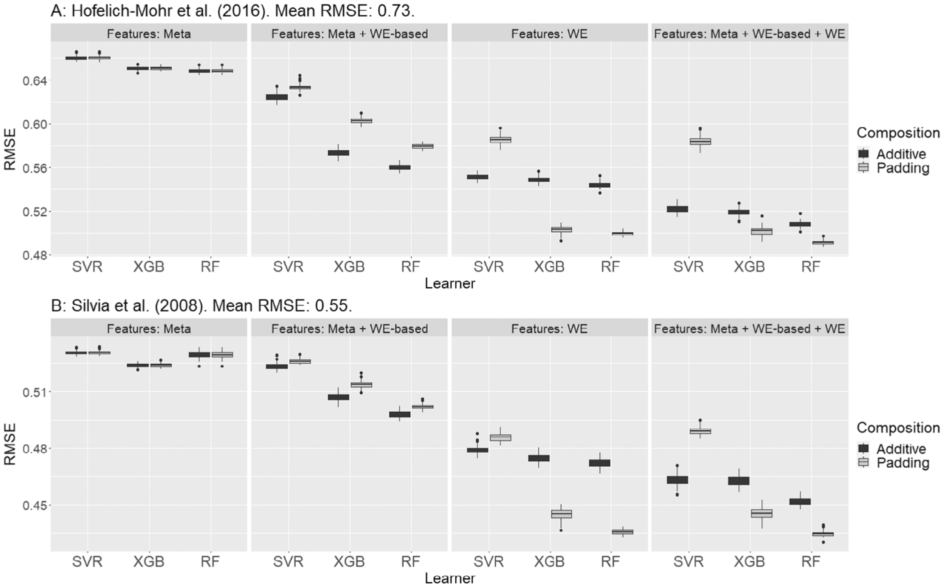

FIGURE A4. Nested resampling RMSE values for additive composition and padding.