Supplementary Material to: Assessing the robustness of automated scoring of divergent thinking tasks with adversarial examples

Yannick Hilker[1], Boris Forthmann[2], and Philipp Doebler[1]

[1]TU Dortmund University

[2]University of Münster

Author Note

Yannick Hilker https://orcid.org/0009-0005-9890-7982, Boris Forthmann https://orcid.org/0000-0001-9755-7304, Philipp Doebler https://orcid.org/0000-0002-2946-8526

Supplementary Material to: Assessing the robustness of automated scoring of divergent thinking tasks with adversarial examples

This Supplementary Material contains details on the experimental evaluation of the adversarial examples based on counter-fitted GloVe word embeddings (Mrkšić et al., 2016; Pennington et al., 2014) instead of the TE3L embeddings in the main manuscript.

Methods

Counter-fitted GloVe word embeddings

GloVe is an unsupervised learning algorithm used to obtain word embeddings, vector representations for various words, which reflect semantic and syntactic relationships between words (Pennington et al., 2014). Each word is assigned one multidimensional vector, so words can be processed computationally. The resulting vector representations have two interesting properties: (1) Semantically similar words are represented by similar vectors. For example, the similarity of the vectors can be determined via the Euclidean distance of the vectors. (2) There is a linear substructure: Other properties are included in the word embedding, such as the relationship between adjectives and their comparatives or superlatives. Word vectors can also be used to represent analogies by performing vector operations. One example is the analogy "king − man + woman = queen".

In the following work, the GloVe word embedding with 400,000 words and 50 dimensions per word is used, trained on Wikipedia2014 and Gigaword text collections. This relatively low dimensionality is sufficient according to Buczak et al. (2023) with more dimensions not leading to noticeable improvements of predictions. In order to use the word embeddings to represent responses, the vectors of the individual words are strung together and padded with zeros so that all vector representations of the answers have the same length.

While GloVe can be used to infer information about the context of different words, sometimes clearly semantically different words are mapped to similar word vectors, such as "north" and "west". To remedy this, Mrkšić et al. (2016) counter-fit the original word

vectors using information from word databases to adjust the original vectors. In the set of the resulting revised GloVe vectors, synonyms are closer together and the vectors of antonyms are further apart. The Mrkšić et al. (2016) word embeddings cover 65,713 words represented in 300 dimensions.

Experimental Evaluation

The experimental evaluation for the counter-fitted GloVe embeddings runs parallel to the evaluation in the main manuscript. The same datasets are used, but the random forest employs the counter-fitted embeddings to predict the human ratings instead of the TE3L embeddings.

Procedure

Also for the counter-fitted GloVe embeddings a random forest with the variance as impurity measure is fitted with hyperparameter tuning (mtry = 49, minimum node size = 5, RMSE = 0.189), with Table 1) containing the accuracy of the predictions on different training and test data sets based on five-fold cross-validation. Similar to the TE3L embeddings, these results are similar to those of Buczak et al. (2023), or are in a similar range between 0.4 and 0.5.

As in the main manuscript, we study the prevalence of adversarial examples, derive a practically relevant threshold value, and describe the increase in robustness after supplementing the training data with the adversarial examples.

Adversarial examples for one word responses

Synonyms and thus adversarial examples are found for 289 of the 364 answers with only one word (79.4% of responses). The distribution of the absolute differences between the prediction of the original word and the prediction of the synonym is slightly skewed to the right (M = 0.372, median = 0.351, SD = 0.225, maximum = 1.197). The maximum difference is obtained by replacing the word “digging” with the word “compass”. In this context, the word “compass” is interpreted as “comprehend”. This example already shows a problem in determining synonyms. Namely, homonyms, words with the same spelling but

different meanings, are interpreted differently depending on the context by humans, but not by the synonym search. The distribution of the differences is slightly left-skewed (M= $0.085 > 0$, median = $0.109 > 0$), indicating that the predictions of the adversarial examples are larger in comparison to the predictions of the original answers.

Table 2 is analogous to Table 2 in the main manuscript, quantifying the deviation of several pairs of variables. The predictions based on the original and adversarial responses have an RMSE of 0.434. In contrast, the RMSE of the prediction of the original data from the true scores is 0.210. This indicates that the deviation of the prediction of the adversarial examples from the prediction of the original responses is higher than the deviation of the prediction of the original responses from the original ratings and thus the deviation of the prediction of the adversarial examples is probably not due to the model variance. An RMSE of 0.565 indicates that the predictions for adversarial examples lead to less precise results. This means that the model is less accurate in predicting adversarial examples than in predicting original data.

Overall, it can be seen that by replacing the individual words with their synonyms, high deviations in the predictions can be achieved in some cases. It should be noted that the synonyms in this case are not restricted by measures such as cosine similarity and thus adversarial examples can arise that may have different meanings than their original. The top 20 adversarial examples are shown in Table 4. When looking at these examples, it is noticeable that some have a significantly different meaning. For example, one person responded “chair” as an alternative use of a brick and the corresponding adversarial example is “electric chair”, which is a synonym but not a plausible alternate use.

Adversarial examples for multi-word responses

Recall, the minimum deviation to flag perturbation as an adversarial example is $min.dif = 0.100$. For answers with multiple words a total of 5,961 adversarial examples are found, based on 1,526 different original responses (65.61% of the 2,326 multi-word responses). The deviation of the model’s prediction on these adversarial examples is shown

in Table 3. We find a mean absolute deviation of 0.212 (SD = 0.109), with adversarial examples having slightly higher predictions of the ratings than the original answers (mean difference = 0.067 > 0), similar to the one word responses.

The error metrics in Table 2 are discussed next for multi-word responses. The deviation of the prediction of the adversarial examples from the prediction of the original responses was greater than the deviation of the prediction of the original responses from the original assessment. Only the MAE increases slightly from 0.175 to 0.209. The RMSE of the prediction of the adversarial examples and the RMSE of the prediction of the original answers in comparison to the original evaluation nevertheless differ (0.436 for the adversarial examples and 0.237 for the original answers), implying poorer predictions in relation to the original evaluation.

In the scatter plot in Figure 1, the prediction of the adversarial examples is plotted on the prediction of the original responses. The blue diagonal is the bisector and values on this line would have the same prediction of the adversarial example and the original assessment. The red lines are parallel to this bisector, but have a distance of 0.33 from it. There are no observations at a distance of 0.1 around the bisector, as adversarial examples that cause a distance of less than or equal to 0.1 are not output by the algorithm. Original responses with low predicted ratings between 1.0 and 1.5 frequently generate adversarial examples with larger predictions. On the other hand, original responses with large predictions between 2.0 and 3.0 generate adversarial examples with lower predictions. As large values are rare in the training data, only very few original predictions exceed 3.0, and none of the few adversarial examples for these have larger predictions than their originals.

Choosing a threshold value

Following the same general strategy as in the main manuscript, Figure 2 shows the trade-off between a substantially different prediction and a sufficient number of examples. We selected a threshold value of 0.2 which is between the RMSE= 0.237 and the MAE = 0.175 of the model and results in approximately 600 adversarial examples and a deviation

of approximately 0.1 of the RMSEs of the prediction of the adversarial examples and the prediction of the original ratings. Multiple altered sentences often achieve no greater deviation than single altered sentences. The predicted scores of multiple permuted sentences often lie between the predicted scores of sentences in which one word was permuted individually.

After fixing the threshold value, the adversarial examples from Algorithm 2 were used in combination with language scores to replace the answers with two synonym perturbations. For this purpose, Algorithm 3 is performed 10 times. After perturbing twice, a total of 241 adversarial examples were found that could not reach the threshold value by replacing only one word. The deviation was improved for a total of 420 out of 1,526 responses (27.5%). In total, 859 adversarial examples are found based on multi-word responses that cause a deviation of more than 0.2.

Description of all adversarial examples

We now summarize the results for the total 1,048 adversarial examples based on all original responses with an absolute deviation of 0.2 or more. This means that adversarial examples were found for 30.7% of the total answers, or 39.0% of the 2,690 answers without duplicates. Figure 3 shows the histogram for the raw differences of the predictions. The random forest tends to predict larger scores.

Table 2 contains the error metrics. The RMSE of the prediction of the adversarial examples compared to the prediction of the original answers is higher than the RMSE of the predictions of the original answers compared to the original evaluation, an increase from 0.281 to 0.385. Comparing the deviation of the prediction of the original answers from the original ratings and the deviation of the prediction of the adversarial examples from the original ratings, a substantial increase of the RMSEs from 0.281 to 0.801 is observed. In summary, the prediction of the adversarial examples is way more error-prone than the prediction of the original answers. This reaffirms that the deviation of the predictions of the adversarial examples from the predictions of the original responses is probably not due

to model variance.

Cosine similarity to the stimulus (3) tends to be higher on average for the adversarial examples (0.302) than for the original responses (0.236). This means that the adversarial examples reflect the context of the stimulus better than the original responses. However, the values of the similarities of the adversarial examples are more dispersed than the originals ($SD_{adversarial} = 0.189$, $SD_{original} = 0.147$), another indication that some adversarial examples do not reflect the context of the stimulus and may therefore be invalid responses to the AUT.

Word stems of the synonyms from the adversarial examples and the words from the original responses are compared next. Therefore, only the adversarial examples for single word responses and the adversarial examples for one perturbation in a response are analyzed. The adversarial examples and the original words are each converted to their word stem and the matches of the word stems are counted. In total, 178 (22.1%) of the analyzed adversarial examples (807) match their synonym in the word stem. In other words, in 22.1% of cases a change in the inflection of the word is sufficient to change the model's prediction.

Retraining with adversarial examples

Similar to the procedure for the TE3L embeddings, we use the adversarial examples are used to train a random forest. 500 adversarial examples are randomly drawn out of the 1,048 adversarial examples and added to the training data set of the random forest model. The results of the five-fold cross-validation of the model are shown in Table 1. The values of the RMSE are similar to the values in Table 1 when cross-validated with the original data. This suggests that the model has adapted to the adversarial examples.

Next, the error metrics of the newly trained model are examined to evaluate how successful the training with adversarial examples is and how robust the model is against adversarial examples after training, see Table 2. Based on the RMSE values, the predictions of the adversarial examples are distinguishable from the predictions of real

data. After training the model with the adversarial examples, the RMSE of the adversarial examples drops to 0.211 in the training data set and to 0.490 in the test data set. The RMSE of the prediction of the training data from the original evaluations is 0.188 and is therefore lower than the RMSE of the prediction of the adversarial examples based on the retrained model from the original evaluations, which is 0.211, indicating a slightly higher uncertainty. However, the adversarial examples in the test data set still lead to significantly higher deviations from the original prediction and the original evaluation. This leads to the conclusion that although the model can adapt to the adversarial examples through targeted training, it is generally unable to predict adversarial examples precisely, as they may be too varied for the model to learn.

Similar to the findings for TE3L, the accuracy of the predictions with the data from Hofelich Mohr et al. (2016) of both the newly trained model and the original model is RMSE = 0.710, which is significantly higher than the RMSEs of the models for the data from Silvia et al. (2008) (RMSE=0.189, 0.188). Buczak et al. (2023) observed a large distributional shift between the mean ratings of Hofelich Mohr et al. (2016) and Silvia et al. (2008), explaining the large differences to some extent. The original model for the data from Hofelich Mohr et al. (2016) performs slightly below the RMSE of the new model when considering predictions of the adversarial examples (RMSE = 0.710, 0.804).

Table 1

Results of the cross-validation (RMSE) on the original data and on the data augmented with adversarial examples.

Data	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Original	0.424	0.444	0.446	0.429	0.434
Augmented data	0.446	0.421	0.475	0.425	0.444

Table 2

Comparison of model performance.

Data	Basis	MSE	RMSE	MAE
single w	Prediction adversarial and prediction Original	0.189	0.434	0.372
	Prediction original and original rating	0.044	0.210	0.145
	Prediction adversarial and original rating	0.320	0.565	0.469
single	Prediction adversarial and prediction original	0.056	0.237	0.209
	Prediction original and original rating	0.056	0.237	0.175
	Prediction adversarial and original rating	0.190	0.436	0.377
multi	Prediction adversarial and prediction original	0.149	0.385	0.358
	Prediction original and original rating	0.079	0.281	0.214
	Prediction adversarial and original rating	0.647	0.804	0.664
augmented	Prediction training and original rating	0.035	0.188	0.126
	Prediction training-adversarial and original rating	0.045	0.211	0.166
	Prediction test-adversarial and original rating	0.240	0.490	0.396

Note: "Single w": single word responses; "single": multiple word responses where only one word is replaced; "multi": multiple word responses where multiple words are replaced; "augmented": training on an extended data set with adversarial examples.

Table 3

Differences of the predictions of adversarial examples and original multi-word responses, and cosine similarity to the stimulus.

Data	Min.	1. Qu.	Median	Mean	3. Qu.	Max.	SD
Differences of the predictions							
raw difference	-0.759	-0.132	0.131	0.067	0.212	0.713	0.229
absolute value	0.100	0.131	0.177	0.212	0.260	0.759	0.109
Cosine similarity to stimulus							
origial response	-0.166	0.134	0.193	0.236	0.311	0.799	0.147
adversarials	-0.249	0.181	0.286	0.302	0.406	1.000	0.189

Table 4

Top 20 adversarial examples for single word responses.

Difference	Previous Word	New Word
1.20	digging	compass
0.93	cut	baseball swing
0.92	buildings	chassis
0.91	house	sign of the zodiac
0.87	kill	belt down
0.85	chimney	lamp chimney
0.84	castle	rook
0.80	sword	brand
0.80	weight	weighting
0.77	stab	knife thrust
-0.65	gift	giving
-0.67	slide	swoop
-0.69	toilet	sewer
-0.70	drum	barrel
-0.72	domino	half mask
-0.85	armor	armour
-0.86	swing	cut
-0.92	knitting	cockle
-0.94	brand	make
-0.98	compass	reach

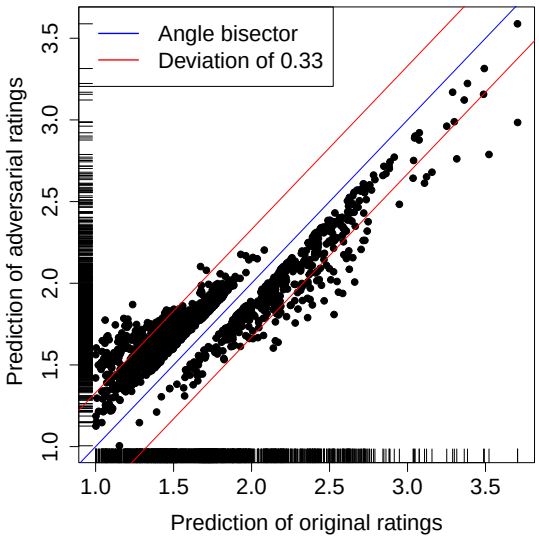

Figure 1

Scatter plot of the predicted ratings for the original data and for the adversarial examples.

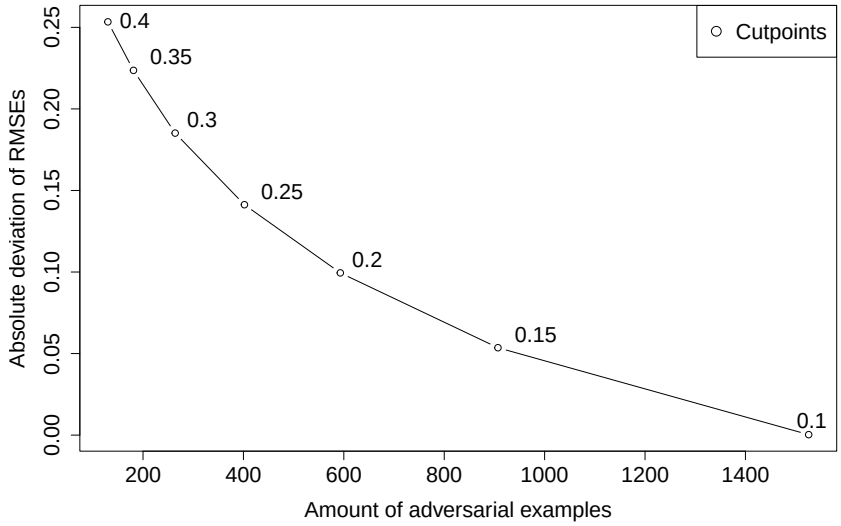

Figure 2

Thresholds related to the number of adversarial examples found and the difference in RMSEs between the predictions of the adversarial examples and the predictions of the original responses and between the predictions of the original responses from the original ratings.

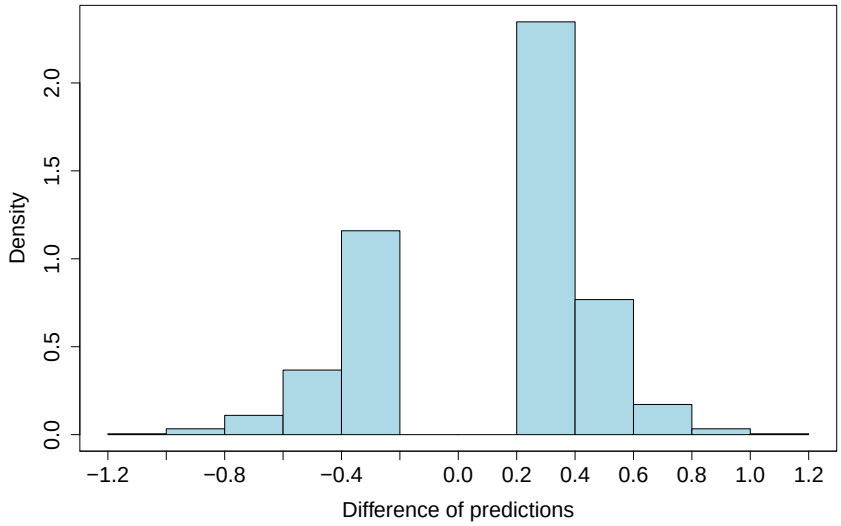

Figure 3

Histogram of the differences between the predictions for the adversarial examples and the original responses.