A Deep Learning Approach to Personality Assessment: Generalizing Across Items and Expanding the Reach of Survey-Based Research

Suhaib Abdurahman[1]*, Huy Vu[2]*, Wanling Zou[3], Lyle Ungar[4], and Sudeep Bhatia[3]

[1]University of Southern California, Department of Psychology

[2]Stony Brook University, Department of Computer Science

[3]University of Pennsylvania, Department of Psychology

[4]University of Pennsylvania, Department of Computer Science

June 24, 2023

https://doi.org/10.1037/pspp0000480

Author Note

Suhaib Abdurahman, Lyle Ungar

All data, analysis code, and research materials are available under https://github.com/goytoom/questionnaire-embeddings. Parts of this work were pre-registered under https://aspredicted.org/dx4dn.pdf. The authors declare no conflicts of interest. Funding was received from the National Science Foundation grant SES-1847794 and a grant from The Honesty Project at Wake Forest University and the John Templeton Foundation.

Correspondence regarding this article should be addressed to Suhaib Abdurahman, Department of Psychology, University of Southern California, SGM 501, 3620 S. McClintock Ave., Los Angeles, CA 90089-1061. Email: sabdurah@usc.edu

Abstract

Traditional methods of personality assessment, and survey-based research in general, cannot make inferences about new items that have not been surveyed previously. This limits the amount of information that can be obtained from a given survey. In this paper, we tackle this problem by leveraging recent advances in statistical natural language processing. Specifically, we extract "embedding" representations of questionnaire items from deep neural networks, trained on large-scale English language data. These embeddings allow us to construct a high dimensional space of items, in which linguistically similar items are located near each other. We combine item embeddings with machine learning algorithms to extrapolate participant ratings of personality items to completely new items that have not been rated by any participants. The accuracy of our approach is on par with incentivized human judges given an identical task, indicating that it predicts ratings of new personality items as accurately as people do. Our approach is also capable of identifying psychological constructs associated with questionnaire items and can accurately cluster items into their constructs based only on their language content. Overall, our results show how representations of linguistic personality descriptors obtained from deep language models can be used to model and predict a large variety of traits, scales, and constructs. In doing so, they showcase a new scalable and cost-effective method for psychological measurement.

Keywords: Personality Prediction; Machine Learning; Natural Language Processing; Measurement

A Deep Learning Approach to Personality Assessment

Personality measurement is central to the study of individual differences and the prediction of behaviors, attitudes, beliefs, and outcomes. Most personality tests use questionnaires composed of natural language items --words and sentences-- that describe common traits. Participants are asked to rate themselves (or others, such as acquaintances) on the trait descriptions, and the resulting data are projected onto a small number of dimensions through statistical techniques like factor analysis. These methods provide quantitative insights into the structure of variability in traits across individuals and are used to motivate influential and highly predictive theories of personality (Digman, 1990; Goldberg et al., 2006; Goldberg, 1990; McCrae & John, 1992).

Despite their successes, traditional methods of personality assessment are constrained to making inferences over the respective set of participants and items in a survey dataset. In other words, factor analysis on standard questionnaire data does not provide any information about the responses of individuals who have not taken the questionnaire, or about participant responses for questionnaire items (and thus traits or constructs) that have not been surveyed. This is particularly relevant for research on understudied populations and traits. Researchers have introduced new tools for addressing the first of these limitations: that of generalizing to out-of-sample individuals. These tools rely on large-scale digital data, such as social media activity, to quantitatively represent thousands of individuals. Researchers give a subset of these individuals a personality questionnaire and, using their responses, build machine learning models capable of predicting the personalities of other individuals using only their digital data (Bleidorn & Hopwood, 2019; Kosinski et al., 2013; Park et al., 2015; Stachl et al., 2021). In this paper we examine whether digital data and machine learning can also provide a solution to the second

limitation: That of generalizing survey data to out-of-sample items, that is, items that have not been surveyed previously (including completely new items that are not part of existing questionnaires).

In order to solve this challenge, we need a way to quantitatively represent the language used in a personality item, so that machine learning models trained on a participant's ratings for one set of items can generalize and make predictions about ratings for a completely new set of items. Recently, a class of deep neural networks, known as transformer models (Devlin et al., 2019; Vaswani et al., 2017) has been shown to accurately represent natural language sentences as *embeddings*. Sentence embeddings (also known as sentence vectors) are points in a high dimensional space, whose structure captures the linguistic properties of sentences. Linguistically similar sentences have similar embeddings and can thus be seen as occupying nearby points in the embedding space. Transformer models are typically trained on large amounts of text data, and the embeddings they extract from this data contain high-quality representations of the linguistic properties of sentences. For this reason, sentence embeddings from transformer models are highly predictive in a variety of natural language processing tasks, including text summarization, sentiment analysis, question answering, and translation (Lewis et al., 2020; Radford et al., 2020; Yang et al., 2019). High quality sentence embeddings are also responsible for the successes of new AI models like GPT3 (Brown et al., 2020), which generate human-like language based on their embedding representations of the user's linguistic input.

One type of transformer model is Sentence-BERT (SBERT) (Reimers et al., 2019). SBERT is specialized for creating embedding representations of sentences that capture their semantics, so that sentences that have similar meanings are given similar embedding representations. SBERT is based on the BERT (Bidirectional Encoder Representations from

Transformers) architecture (Devlin et al., 2019; Liu et al., 2019), and is trained on a very large language dataset, as well as a large dataset of sentence pairs annotated with linguistic entailment relations (Bowman et al., 2015; Cer et al., 2017). Subsequently, embeddings produced by SBERT have outperformed many other methods in the SentEval evaluation set of tasks (Conneau et al., 2018). We take advantage of this model in our study and use SBERT to extract embeddings for questionnaire items (Figure 1A). This allows us to describe personality item sentences as points in a high-dimensional semantic space (Figure 1B), in which items with similar meanings are located close to each other. Since SBERT can be used to obtain embedding representations for *any possible personality item* we can use it to generalize from a small set of rated items to thousands of new unrated personality items. This can be accomplished using several standard machine learning techniques that use the similarities between representations for generalization. The K-Nearest Neighbors regression (Cover & Hart, 1967), for example, predicts the rating of a new item by averaging the ratings assigned to the *K* nearest items in the embedding space (Figure 1B).

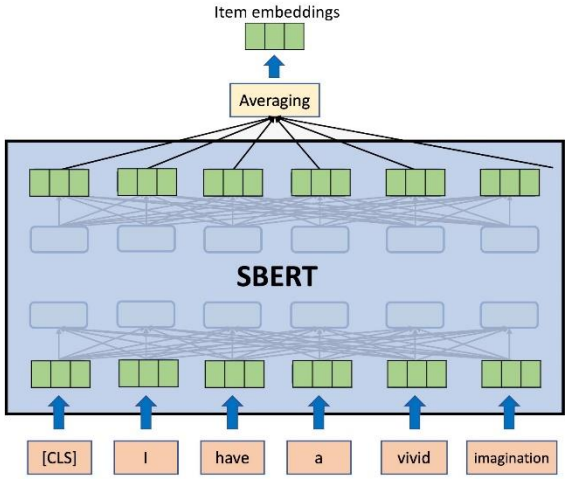
Theoretically, our approach draws on the lexical hypothesis in personality psychology (Allport & Odbert, 1936; Cattel, 1943; Galton, 1884), which proposes that personality traits that are important to a group are expressed through words and sentences in their language. This hypothesis has motivated many advances in personality research and is the foundation for leading personality theories, such as the five-factor model, which uses natural language to describe and measure the core dimensions of variation in people's personalities (Goldberg, 1990; John et al., 1988; Klages, 1929). The lexical hypothesis also implies that quantitative representations of language obtained from deep neural networks should make it possible to quantify, and subsequently predict personality traits, since these models are based on the

statistics of everyday language. In other words, traits that co-occur with each other should have similar linguistic descriptors, and deep networks trained to quantify these descriptors should be able to generate similar representations for the traits.

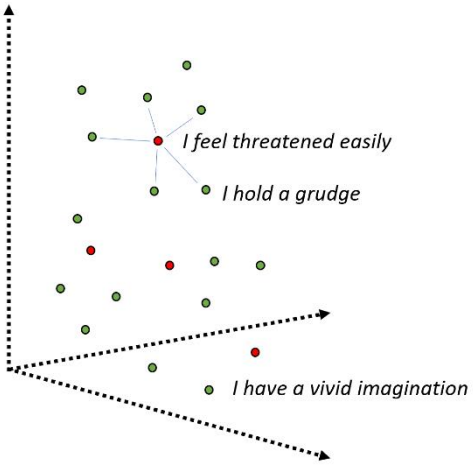
Figure 1

Overview of the Modeling Approach

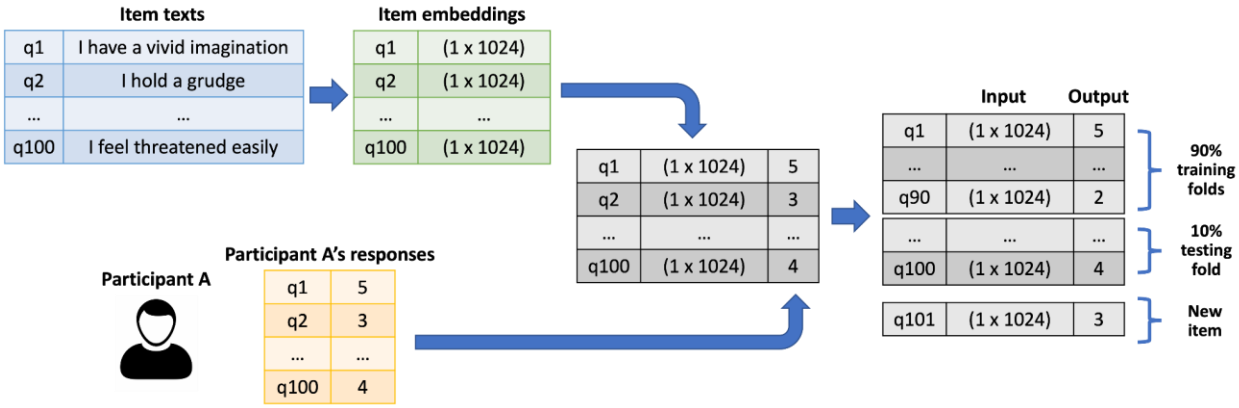
A.



B.



C.



Note. Panel A: Illustration of SBERT, which takes sentences (personality items) as inputs and generates 1,024 dimensional embeddings (vectors) of the sentences as outputs. B: Illustration of the SBERT embedding space and K-Nearest Neighbors regression. In this space, each point represents an item, and items with similar meanings are located close to each other. Predictions for test items (indicated in red) can be obtained by averaging the ratings of

the *K*, here 5, nearest training items (indicated in green). Note that the actual SBERT embedding space is 1,024 dimensional instead of 3-dimensional, and that any possible personality item (if expressed as a sentence) can be represented as a point in this space. C: Methodology for predicting participants' responses from item embeddings in Study 1. We first extract SBERT embeddings for all items in the NEO-PI-R questionnaire (100 items in this example). Then, for each participant, we use these 1,024-dimensional item embeddings as inputs and the participant's responses as target outputs in a standard machine learning model, such as K-Nearest Neighbors regression. We test the model's performance using 10-fold cross-validation. The fully trained model can then be used to predict the participant's responses to new items that are not in the questionnaire.

Our approach is also inspired by recent work that has used the semantic similarity of the individual words in questionnaire items to measure the similarity of items (Arnulf et al., 2014; Evans et al., 2022; Garcia et al., 2020; Rosenbusch et al., 2020). This line of research leverages word embeddings (high dimensional vectors for individual words) for item selection, e.g., finding related items for a new scale or avoiding redundancies. We extend this idea to predict participant responses, a problem that can now be solved as deep language models provide high-quality representations for sentences that are based not just on the individual words in the items, but also syntax and word order in the sentence (in this way sentence embeddings capture nuances in sentence meaning that cannot be captured by word embedding models). In our analysis below, we also consider a version of our approach applied to word embeddings, a method similar to that of Rosenbusch et al. (2020), to test whether sentence embeddings from deep neural networks provide superior representations and predictions for personality items.

We evaluate the applicability of our approach for personality prediction using a series of empirical tests. In Study 1, we apply cross-validation to a 100-item NEO-PI-R dataset to assess our model's ability to predict out-of-sample; that is, we train our model on ratings for a subset of items and use it to predict ratings for held-out items (Figure 1C). We also contrast our models' accuracy rates with those of human judges that are given an identical task and are incentivized to

make accurate predictions. To ensure that our approach is robust, we replicate this analysis pipeline for three additional personality questionnaires in Study 2A-C. In Study 3, we extend our tests to a large new dataset of over three thousand existing personality items taken from hundreds of different scales and constructs, thus allowing us to test the cross-domain predictive accuracy of our approach. Finally, in Study 4, we use SBERT's assessment of item similarity to infer the personality dimensions and constructs associated with personality items. This allows us to test whether established constructs are explicitly reflected in the structure of the underlying SBERT embedding space. We also attempt to cluster novel, unlabeled personality items with already established and labeled ones, in order to infer personality constructs associated with these items.

If successful, the methods outlined in this paper would provide a powerful set of tools for personality researchers. By predicting responses of participants to thousands of personality items, researchers would be able to describe each individual in terms of a large variety of distinct traits, scales, and constructs, and in turn build richer models of personality (and associated behaviors, attitudes, beliefs, and outcomes), without the need for extensive data collection. Additionally, new personality items could be tested using our trained models instead of human participants. This would offer researchers a cost-effective and scalable resource for psychological measurement and theory development (Bleidorn & Hopwood, 2019; Evans et al., 2022; Yarkoni & Westfall, 2017).

Transparency and Openness Promotion.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow JARS (Appelbaum et al., 2018). All data, analysis code, and research materials are available at https://osf.io/sxg8n/?view_only=762b6188d54246c0a4c1c7e6218e33c3. Data were analyzed

using Python, version 3.9 (Van Rossum & Drake, 2009) and the packages sklearn, version 0.23 (Pedregosa et al., 2011), statsmodel, version 0.13 (Seabold & Perktold, 2010), and matplotlib, version 3.5.0 (Hunter, 2007). Study 1's depth analysis experiment was pre-registered on AsPredicted.org [masked copy available in the supplemental materials].

Ethics Statement

All studies involving human data collection were approved by the Institutional Review Board of [masked affiliation, provided after acceptance] under approval number [masked; provided after acceptance]. Computational studies without human data or with only retrospective analysis of anonymized human data from public data sets were exempted from the need for approval.

Computational Methods

We begin by summarizing the computational methods used in our studies. These methods entail the generation of questionnaire items' embeddings (i.e., SBERT and baselines), training, validation, and selection of the predictive models, as well as evaluations of model performance. Any deviations from and extensions of these methods are detailed in the respective studies' sections.

SBERT Embeddings

We created the questionnaire item embeddings by feeding each item's text through the SBERT neural network, then averaging the embeddings in the last layer across the whole sentence to get a representation for the item. See Figure 1A for a visual overview of this procedure. We used the SBERT version called "nli-roberta-large", based on the pre-trained RoBERTa-Large model (Liu et al., 2019) with 24 layers and 1024 hidden vector dimensions, and trained on the MultiGenre NLI dataset (Williams et al., 2018). RoBERTa is based on the BERT

(Devlin et al., 2019) architecture, and outputs a 1024-dimensional embedding for each of the questionnaire items. For the implementation of the embedding extraction, we used the Python library sentence-transformers, provided by the original authors (Reimers et al., 2019). Note that the item responses for the personality prediction task were not reversed-coded and no information about items' construct or direction they load onto a construct (positive/negative) was provided to the model, meaning our model based its representations solely on the items' raw text.

Baseline Embeddings

We considered two alternative embeddings as baselines to SBERT, Word2Vec (Mikolov et al., 2013) and Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015). Word2Vec was trained on English Wikipedia texts and generates 300-dimensional embeddings for millions of common English words. These embeddings have been shown to accurately capture the similarity relationships between individual words, so that similar words have similar word embeddings (Richie & Bhatia, 2021). To generate Word2Vec embeddings for each item, we averaged the word embeddings across all words in the item text. Importantly, Word2Vec is only sensitive to the individual words from the personality item, and, unlike SBERT, ignores the ordering of the words and syntactical structure of the sentence. Therefore, this model is similar to prior attempts at measuring item similarities using the similarities of their individual words (Arnulf et al., 2014; Evans et al., 2022; Garcia et al., 2020; Rosenbusch et al., 2020).

LIWC provides sets of words (called "lexicon") for 73 common psychological variables. For example, the Anxiety variable contains words like "worried", "fearful", and so on. To obtain LIWC embeddings for our personality items we simply counted the number of times words in each of the 73 LIWC lexicons occurred in the personality item, giving us 73-dimensional LIWC

embeddings. The counting and scoring of personality items on LIWC dimensions were done using the Differential Language Analysis ToolKit (Schwartz et al., 2017).

Predictive Models

We applied standard machine learning techniques, such as Ridge (Hoerl & Kennard, 1970), K-Nearest Neighbor (K-NN; Cover & Hart, 1967), and Support Vector (SVC; Hearst et al., 1998) classification and regression, to map each item's embeddings onto each participant's rating of the item. Ridge methods made predictions by estimating a regularized linear function on each embedding dimension, K-NN methods made predictions by averaging the ratings on the K most similar items to a target item and SVC methods made predictions by estimating a potentially nonlinear function on the embedding dimensions with a kernel trick (i.e., mapping the input into a further high-dimensional space).

We applied the regression and classification methods to all three types of embeddings: SBERT, Word2Vec, and LIWC. This gave us a total of 5x3 = 15 models (e.g., KNN regression with SBERT embeddings, SVC with LIWC embeddings, etc.). Regarding hyperparameter-tuning, for the Ridge models we tested five different *a*-values, corresponding to the weight on the regularization penalty. For the KNN methods we tested five different *K* values, reflecting the number of nearest neighbors used for prediction. For the SVC method we tested five different *C* values, corresponding to the strength of the regularization penalty. To ensure consistency across our studies, we determined the best performing embedding and hyperparameter combination in Study 1 and used it in all subsequent studies. See Table A1 in the Appendix for the specific values used and their effect on performance in Study 1.

Cross Validation

We trained and evaluated models using 10-fold cross validation. In particular, we divided each participant's data into ten equally sized groups or folds (with 10% of ratings in each fold), then fitted each model on the 90% of items in the first nine folds (the training data) and evaluated its predictions on the ten items in the held-out fold (the test data). This was repeated nine more times with each fold serving as the test data once. See Figure 1C for an illustration. All models were estimated in the Python scikit-learn library (Pedregosa et al., 2011).

To evaluate model performance, we calculated the correlation of a model's predictions with the observed ratings for each participant. Specifically, for the *i*th participant and the *k*th test-fold ($k = 1, 2 \ldots, 10$), we calculated a model's predictions for the ten items in the fold, when that fold was in the test data. We then concatenated the predictions across all testing folds into one list containing our model's out-of-sample predictions for each item offered to the participant. This list was then compared with the observed ratings of the participant using Pearson correlation, to obtain a measure of our model's accuracy for that participant.

Study 1

In Study 1, we tested whether our approach, as described in the Computational Methods section, accurately predicts participant responses to out-of-sample items in an established personality questionnaire: the NEO-PI-R (Costa & McCrae, 1992; Goldberg et al., 2006). We also evaluated our approach against previous models in the field and against incentivized human judges.

Methods

NEO-PI-R Dataset. We used data collected by Stillwell & Kosinski (2012) in order to train and test our predictive model. This dataset has responses from $N = 2{,}749$ individuals who

used the myPersonality Facebook application between 2007 and 2012. Personality was measured using the NEO-PI-R five-factor model (Costa & McCrae, 1992), and for this reason we will refer to this dataset as the *NEO-PI-R dataset* for the rest of the paper. The five-factor model classifies each participant along five personality dimensions: Openness to experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). The questionnaire contains 100 natural language items (20 per personality dimension) from the International Personality Item Pool (IPIP; Goldberg et al., 2006). Each item asks participants to indicate their agreement with a description on a five-point Likert scale. Participants in the NEO-PI-R dataset completed all 100 items. See Table 1 for a summary of dataset characteristics.

Table 1

Key Characteristics of Each Questionnaire Used in This Paper

	# Items	# Participants	# Constructs	Reverse Coded
Study 1 NEO-PI-R	100	2,749	5	Yes
Study 2A 16PF	163	49,159	16	Yes
Study 2B RIASEC	48	135,764	6	No
Study 2C HSQ	32	590	4	Yes
Study 3 IPIP	3,653	161	242*	Yes*

Note. The Reverse Coded column refers to whether the questionnaire involves reverse coded items that have negative loadings onto the construct. For the IPIP, information regarding constructs and directions was only available for 1,932 items out of 3,653 items.

Comparison with Human Judges. We compared our model's performance with that of incentivized humans that were given an identical prediction task (i.e., predicting a participant's

NEO-PI-R ratings on 10 test items using their ratings on 90 training items). There were 2,749 participants and 10 cross-validation folds per participant in the NEO-PI-R dataset, generating 27,490 distinct prediction tasks for our model. Since collecting human predictions for so many tasks was not feasible, we compared model vs. human performance on a subset of participants from the original study. Specifically, we conducted two human rating experiments which prioritized depth (i.e., a detailed analysis of target-level human accuracy that reduces noise on the target-level estimates) and breadth (i.e., covering a broad range of model performance to maximize representativeness) respectively.

Depth experiment. For our first experiment, we selected three target participants out of the 2,749 participants from the NEO-PI-R dataset based on the predictive accuracy of our main SBERT model (details in results section below). Specifically, we selected one target participant for whom our model performed well (75th percentile accuracy out of all participants), one participant for whom our model performed moderately (50th percentile accuracy), and one for whom our model performed poorly (25th percentile accuracy). The 100 responses of each of these three target participants were divided into the same 10 folds as used in the cross-validation analysis for model training and evaluation, resulting in a total of 30 tasks for our human judges.

We collected a sample of 600 human judges (41.83% female; Mean age = 36.17) from Prolific Academic for this part of Study 1. Each human judge was randomly assigned to one task leading to 20 judges per task, reducing noise and allowing for rigorous tests of target-level human accuracy. In the training phase of the task, the human judge first viewed a target participant's responses to 90 NEO-PI-R survey questions one at a time, and then in the testing phase, they predicted the target participant’s responses to the remaining 10 NEO-PI-R survey questions in the testing fold. Each participant received $2 for completion. We also incentivized

participants by giving a bonus payment of $1 to those whose predictive accuracy was in the top 10% among all judges. We preregistered this study at AsPredicted.org [masked copy available in the supplemental materials].

Breadth experiment. Our second experiment obtained human ratings for 60 target NEO-PI-R participants. We chose the target participants corresponding to the 0^{th} to 100^{th} percentiles of model predictive accuracy in equidistant steps (e.g., 0^{th}, 1.5^{th}, 3^{rd}, 4.5^{th} percentile accuracy, and so on). This way, our sample of target participants covered a broad range of our model's predictive accuracy allowing for a more representative comparison of the model and human performance (i.e., comparing human predictions to several bad, average, and good model predictions). We collected a sample of 600 human judges (45.5% female; Mean age = 38.85) from Prolific Academic. All other aspects of the design were identical to the depth experiment, except for the fact that there were 600 tasks (60 target participants X 10 cross-validation folds) leading to only one human judge per task.

Results

Model Performance. We found the best performing model using SBERT embeddings to be the K-NN regression with $K = 5$ (see Table A1 for a full model comparison). Intuitively, this model finds the $K = 5$ training items that have the most similar embeddings to the test item and then averages the participant's rating on these items to predict their rating of the test item. For Word2Vec and LIWC embeddings, we found the best performing model to be SVC with $C = 10$ and Ridge regression with $a = 10$ respectively.

The best-performing SBERT model achieved an average correlation of .45 in predicting out-of-sample ratings (one-sample t-test against zero mean: $t(2{,}747) = 158.09$, $p < .001$; 95%-CI .45 to .46). As can be seen in Figure 2A, the models with the Word2Vec and LIWC embeddings

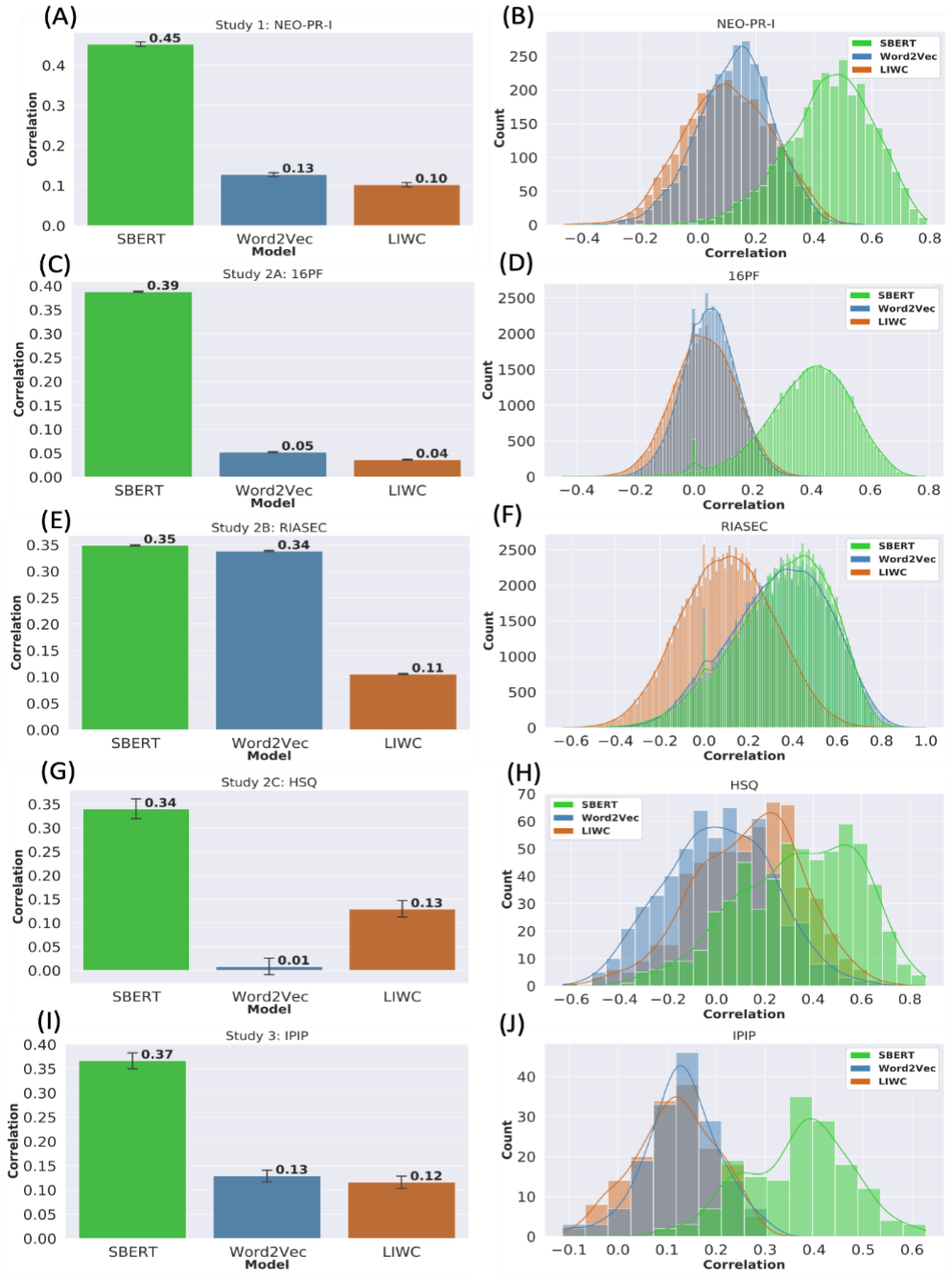
achieved much lower performances than the model with the SBERT embeddings, with average correlations of .13 ($t(2,747) = 51.68$, $p < .001$; 95%-CI .12 to .13) and .10 ($t(2,747) = 35.43$, $p < .001$; 95%-CI .01 to .11) respectively. This indicates SBERT's superiority in quantifying item representations. In Figure 2B we present the distribution of correlations across participants. Here we can see that the SBERT model achieved a significantly ($p < .05$) positive correlation for the vast majority (94.2%) of participants. By contrast, the Word2Vec and LIWC models achieved significant positive correlations for a much smaller proportion of participants (30.2% and 28.0% respectively).

Comparison with Human Judges. To contrast our model's performance against human judges, we calculated the judges' performance as the correlation between the predicted item responses and the observed item responses for each test fold on each target participant. Specifically, for the i^{th} target participant ($i = 1, 2, 3$), and the k^{th} fold ($k = 1, 2 \ldots, 10$), we calculated the human prediction for the ten items in the fold. These ten predictions were then compared with the observed ratings of the target participant to generate a fold-level Pearson's correlation. We calculated the K-NN SBERT model's performance analogously.

Figure 3A shows that our method's performance, in our depth experiment, was on par with that of human judges. The average correlation across 10 testing folds of our model for the 75^{th} percentile target was .61 (one-sample t-test against zero-mean: $t(9) = 9.75$, $p < .001$; 95%-CI .47 to .75), 50^{th} percentile target was .46 ($t(9) = 4.86$, $p < .001$; 95%-CI .24 to .67), and 25^{th} percentile target was .38 ($t(9) = 4.07$, $p = .003$; 95%-CI .17 to .59). Meanwhile, the average correlation of 200 human judges for the 75^{th} percentile target was .43 ($t(199) = 17.51$, $p < .001$; 95%-CI .38 to .47), for the 50^{th} percentile target was .58 ($t(199) = 31.54$, $p < .001$; 95%-CI .54 to .62), and for the 25^{th} percentile target was .39 ($t(199) = 17.8$, $p < .001$; 95%-CI .34 to .43).

Figure 2

Prediction Accuracy of the K-NN SBERT Model vs Baselines for All Datasets

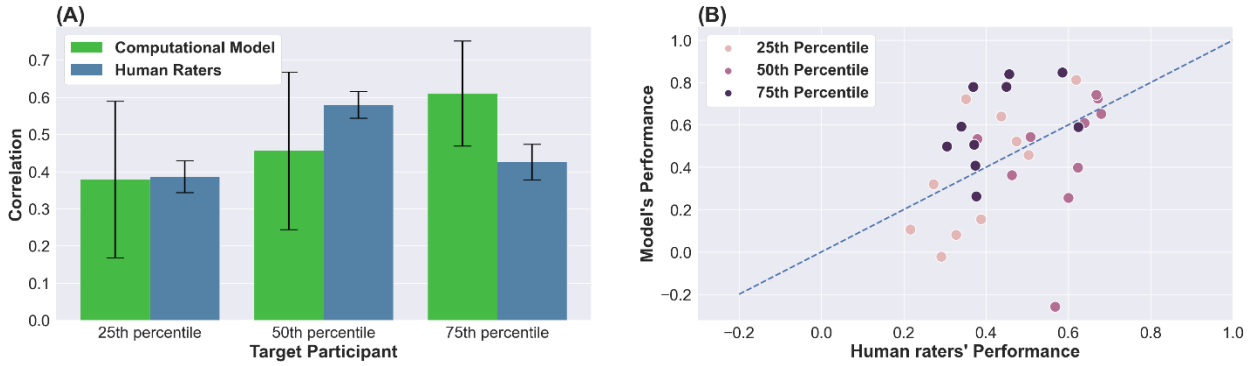


Note. We measured prediction accuracy of each model by calculating the correlation between the model's predicted and observed out-of-sample responses for each participant. The average correlation across all participants is reported in the left panel of each questionnaire (error bars correspond to 95%-confidence intervals for the average correlation). The distribution of correlations across all participants is reported in the right panel.

In total, the average correlation across all three targets was slightly better for our model: .48 ($t(29) = 9.57$, $p < .001$, 95%-CI .38 to .59) vs. .47 ($t(599) = 36.09$, $p < .001$, 95%-CI .44 to .49), though not statistically distinguishable (Welch's two-sample t-test, independent samples: $t(32.92) = -0.32$, $p = .751$). Notice that the error bars were larger for our models relative to the human judges as its correlations were averaged across only 10 values (one value for each of 10 testing folds), instead of 200 values for human judges (20 judges for each of 10 testing folds).

Figure 3

Prediction Accuracy of the K-NN SBERT Model vs Human Judges



Note. The performances of the two were tested on the personality ratings of three target participants. Panel (A) shows the average correlations between predictions and observed responses across experimental conditions. 95%-confidence intervals are reported as error bars. For each target participant, the reported correlation was averaged across the $N = 200$ human rating correlations and $N = 10$ model prediction correlations. Panel (B) shows correlations of predictions and observed responses for each of $N = 30$ experimental conditions. Each point in the plot refers to the performance of our model (y-axis) and the average human judge (x-axis). The points were grouped by target participant (color). The plot shows that our model performance was comparable to that of human judges.

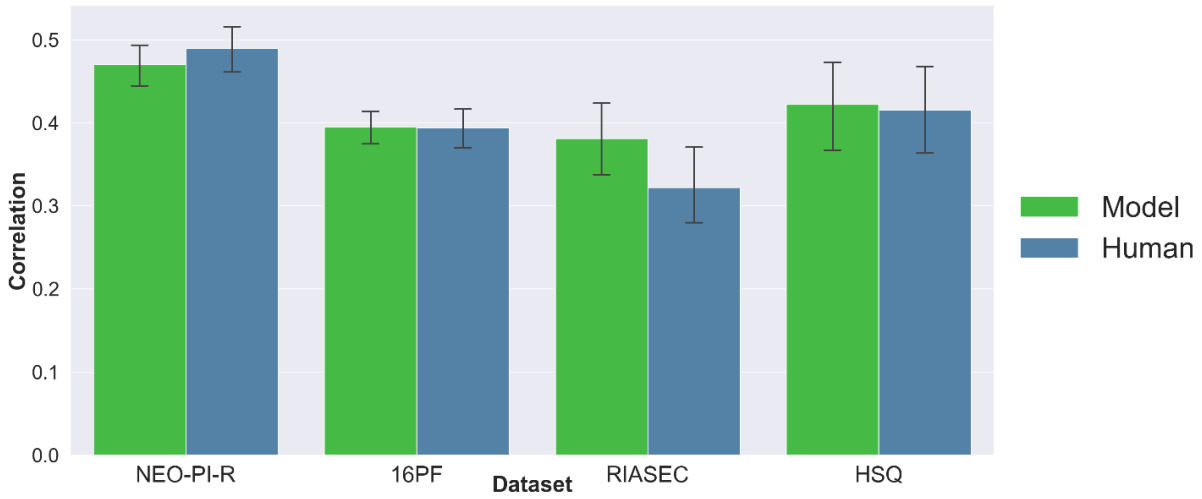
Figure 3B shows the SBERT model's performance and the human judges' performance for each of the 30 tasks in the depth experiment (10 distinct testing folds, for 3 target participants). Here we see that the model performance was roughly equal to the human performance across the tasks. Note that there was one outlier condition for the 50th percentile

target, for which our model achieved a correlation of -.26 ($r(8)$, $p = .47$) whereas human judges achieved an average correlation of .57 (one-sample t-test against zero mean: $t(19) = 10.53$, $p < .001$; 95%-CI .46 to .68). The outliers seem to be driven by contradictory statements in the conscientiousness training items (high and low), which led to neutral predictions of these items. It could be that the human judges were less sensitive to these contradictions because they ignore contradicting facets (e.g., using only the most relevant facet) or use their expectations regarding social faking. This explains the slight underperformance of our model for the second (50th percentile) target.

Regarding the second, breadth-focused experiment, Figure 4 shows that our method's performance was, again, on par with that of human judges and that the overall distribution of prediction performance across experimental conditions was similar. Specifically, the average correlation across all experimental conditions was .47 ($t(598) = 38.39$, $p < .001$) for our model, and .49 ($t(598) = 36.375$, $p < .001$) for the human judges. The difference between model and human performance was statistically non-significant (Welch's two-sample t-test, independent samples: $t(1183.06) = 1.08$, $p = .28$).

Figure 4

Comparison of our method with incentivized human judges

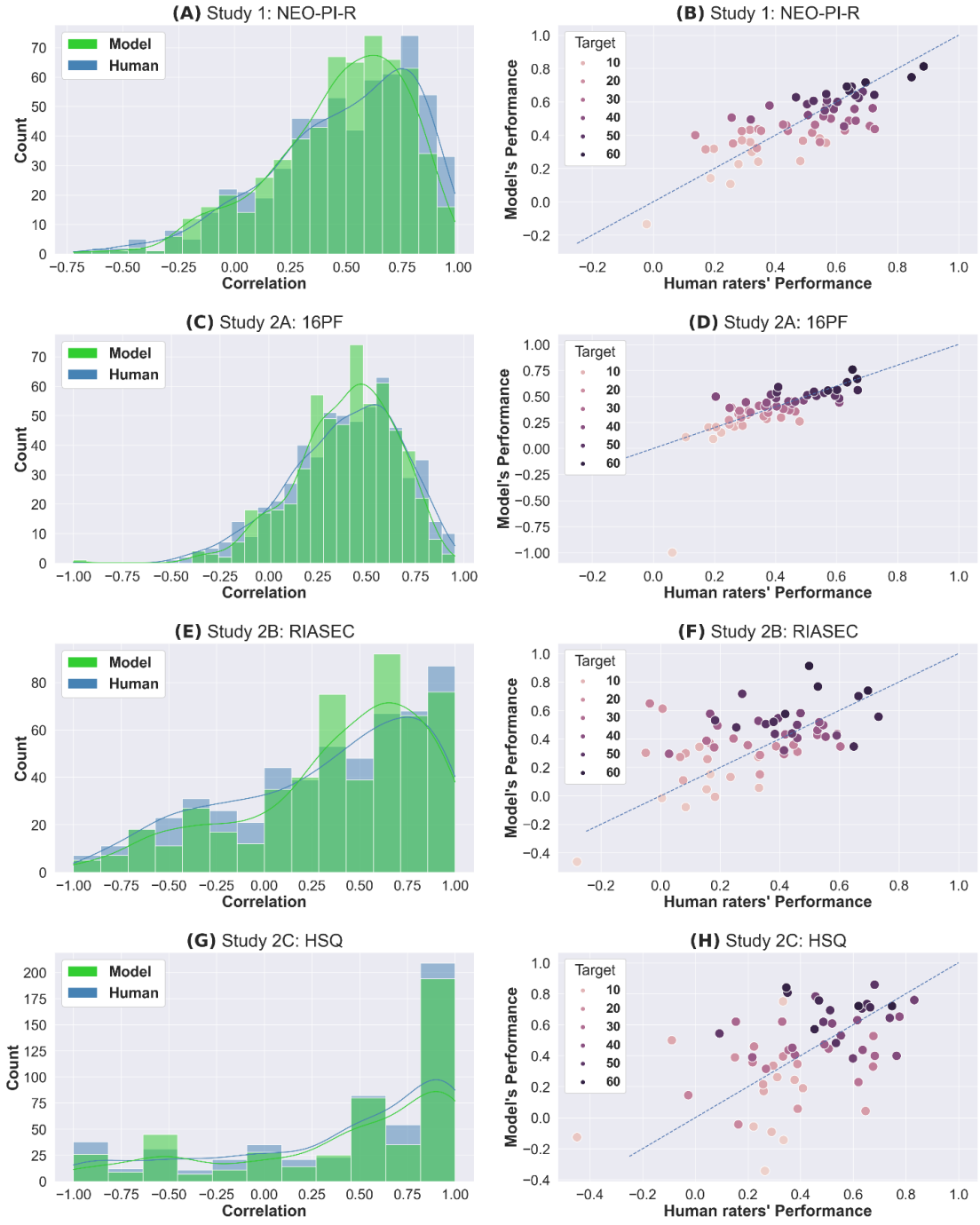


Note. The performances of the human judges and our model were tested on the personality ratings of 60 targets from the original questionnaire datasets. Plot shows the average correlations between predictions and observed responses across all experimental conditions with 95%-confidence intervals as error bars. The figure shows that our model's performance is on par with incentivized human judges across all questionnaires.

Importantly, Figures 5A and 5B show that the performances of both our model and the human judges were not driven by outliers. For example, Figure 5A shows that most of the model and human performance is positive (> 90%), with the majority being larger than .50. Figure 5B shows that the performances on most targets are very similar (close to dashed line/equality). The figure further shows that model and human performance strongly correlated, $r(58) = .76$; $p < .001$, indicating that the prediction problems were similar for both model and humans (i.e., targets that were easy to predict for our model were also easy to predict for humans).

Figure 5

Detailed performance overview of our model and incentivized human judges



Note. Detailed analysis of our model's and human judges' performance. The left column shows the distribution of prediction accuracy -- correlation between predicted and observed personality ratings -- across all experimental conditions for both human judges and our model. The distribution plots indicate similar performance for our model

and human judges. The right column shows correlations of predictions and observed responses for each of the 60 targets (averaged across all test folds). Each point refers to the performance of our model (y-axis) and the average human judge (x-axis) for a specific target and test fold. Dashed line represents equality. The figures show that our model performance was comparable to that of human judges (close to dashed line) and that neither performance was driven by outliers.

Discussion

Study 1 tested our approach to predicting participant responses on out-of-sample personality items, using a large existing dataset of NEO-PI-R ratings. It found that sentence embeddings obtained from SBERT (a leading deep language model optimized for encoding semantic similarity between sentences) were able to predict out-of-sample participant responses with an average correlation of .45. To interpret our accuracy rates, we can contrast our results with previous predictive models of five-factor responses. Liu et al. (2016) achieved a maximum correlation of .19 using social media profile picture features; Golbeck (2016) reported a correlation of up to .24 using texts from participants' public Facebook posts; and Youyou et al. (2015) showed an average correlation of up to .56 using a large amount of a person's Facebook likes. Our approach differed from these tests in two ways. First, we predicted responses for out-of-sample items, whereas all prior tests predicted responses for out-of-sample individuals. Second, we predicted responses at the item level, instead of the average personality scores on a construct (e.g., an Openness score). There is much higher variability in responses on the item level, making our prediction problem significantly harder. Finally, it should be noted that the performance metrics for Liu et al. (2016) and Golbeck (2016) reported above, were the metrics for the best performing constructs (e.g., Agreeableness for Golbeck (2016)) and not the average performance across all items as reported for our method. For this reason, we can conclude that our prediction exercise was quite successful, especially relative to past work on personality prediction.

We also tested various machine learning models for predicting responses from SBERT embeddings. Here we found that a K-Nearest Neighbors regression with $K = 5$ performed the best. This model predicts the rating for an out-of-sample item by finding the five training items with the highest similarity and then averaging the participant's ratings for those items. The superiority of the K-NN algorithm over alternate models like the Ridge regression indicates that the mapping between item embeddings and participant ratings may be non-linear. K-NN can successfully capture these non-linearities, as it uses only the local structure of input space (five nearest neighbors) to predict responses.

We also found that SBERT provided better item representations, and subsequently higher predictive power, than embeddings obtained from Word2Vec (a prominent word representation model) and LIWC (a common approach to extracting psychological variables from text). The superiority of SBERT over Word2Vec and LIWC indicates that the context and sentence structure (e.g., word order, syntax) of a personality item plays an important role in specifying its meaning, and that averaging embeddings for the words in the item (as with Word2Vec) or counting up the words associated with various psychological variables (as with LIWC) is not enough for capturing the psychological richness of the item. Prior work has used word embeddings for scale creation, e.g., tasks such as item selection (Arnulf et al., 2014; Evans et al., 2022; Garcia et al., 2020; Rosenbusch et al., 2020). Our results show that it may be possible to improve on this work by using SBERT embeddings to measure item similarity.

Finally, and perhaps most importantly, we ran two experiments in which we elicited incentivized personality predictions from human judges. The first of these was a depth experiment, which obtained a large number of human predictions for each cross-validation task performed by our model but used only three NEO-PI-R target participants. The second was a

breadth experiment, that used a much larger set of NEO-PI-R participants but obtained a smaller number of human predictions per cross-validation task. Both these experiments showed that our model performed equivalently to human judges, indicating that it is as good at out-of-sample personality prediction as humans. Additionally, we found that human and model predictions were highly correlated with each other (r = .76) indicating that target participants that were easy or difficult for humans to predict were also easy or difficult for the model to predict. This provides strong evidence for the capabilities of our model, and in particular the quality of its representations for the items that make up the NEO-PI-R survey.

Studies 2A-C

Overall, Study 1 shows that deep language models provide high quality quantitative representations for NEO-PI-R personality items, and that the similarities between these representations can be used to predict responses for out-of-sample items with a human level of accuracy. In Studies 2A-C, we tested whether our model's performance generalizes beyond the NEO-PI-R to three other personality questionnaires: 16PF (16 Personality Factors; Cattell & Mead, 2008) in Study 2A, RIASEC (Realistic, Investigative, Artistic, Social, Enterprising, Liao et al., 2008) in Study 2B, and HSQ (Humor Styles Questionnaire, Martin et al., 2003) in Study 2C. This diverse conceptualization and structure of personality allowed us to test the generalizability of our model.

Methods

Datasets. We used data collected from the Open-Source Psychometrics Project (https://openpsychometrics.org/_rawdata/) for the 16PF questionnaire (N = 49,159 participants), RIASEC questionnaire (N = 135,764 participants), and HSQ questionnaire (N = 590 participants). These questionnaires had 163, 48 and 32 items respectively, and each participant

rated each item on a five-point Likert scale. See Table 1 for a summary. We selected these three datasets (and not others in the Open-Source Psychometrics Project), as they are all multi-construct Likert-scale questionnaires with at least 500 responses, covering diverse topics and involving diverse constructs. For instance, the 16PF questionnaire is another hierarchical personality model based on the lexical hypothesis (Rossier et al., 2004), similar to the NEO-PI-R. However, the 16PF hierarchical structures were designed using a bottom-up structure (identifying 16 primary factors and then five higher level dimensions) as opposed to the NEO-PI-R's top-down approach (identifying five higher level dimensions and then 30 lower-level facets). The 16PF questionnaire contains the following constructs: Warmth (A), Reasoning (B), Emotional stability (C), Dominance (E), Liveliness (F), Rule-consciousness (G), Social boldness (H), Sensitivity (I), Vigilance (L), Abstractedness (M), Privateness (N), Apprehension (O), Openness to change (Q1), Self-reliance (Q2), Perfectionism (Q3), and Tension (Q4) (Rossier et al., 2004). The RIASEC questionnaire describes personality through preferences and aversions that influence the choice of work environments (and environments through typical work activities and demands placed on individuals). The questionnaire contains six personality dimensions (and parallel environments): Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C), collectively called RIASEC (Armstrong et al., 2008). Finally, the HSQ describes personality through different styles of using humor, containing the dimensions Self-enhancing, Affiliative, Aggressive, and Self-defeating (Martin et al., 2003). As such, the HSQ uses a conceptualization of humor as a stable multi-dimensional aspect of personality (Lopez & Snyder, 2003). Validation studies have further shown that the HSQ dimensions (humor styles) correlate with other established personality measures, such as the NEO-PI-R dimensions (Martin et al., 2003).

Comparison with Human Judges. Analogous to Study 1, we compared our model's performances on the above questionnaires with that of human judges incentivized to make accurate predictions. We collected a sample of 600 judges for each of the questionnaires (16PF: 62.00% female, Mean age = 39.71; RIASEC: 51.33% female, Mean age = 42.41; HSQ: 50.83% female, Mean age = 40.60). All test procedures were the same as the breadth experiment in Study 1. In particular, human judges predicted the responses of 60 target participants from the original questionnaire datasets (i.e., target participants for which our model performance ranged from the 0^{th} percentile to 100^{th} percentile). Each judge was given one of the model's cross-validation tasks, in which they used 90% of the target's responses (training fold) to predict the held out 10% of their responses (test fold). We incentivized participants by giving a bonus payment of $1 to those whose predictive accuracy was in the top 10% among all judges.

Results and Discussion

Figure 2C-H show that the out-of-sample predictive performance of the SBERT method persisted for these questionnaires, with average correlations of .39 ($t(49,158) = 585.49$, $p < .001$; 95%-CI .39 to .39) for 16PF in Study 2A, .35 ($t(135,763) = 594.91$, $p < .001$; 95%-CI .35 to .35) for RIASEC in Study 2B, and .34 ($t(589) = 33.21$, $p < .001$, 95%-CI .32 to .36) for HSQ in Study 2C. Overall, our SBERT model achieved significant ($p < .05$) positive correlations for 93.3%, 51.4%, and 64.7% of participants in the three studies respectively. Consistent with Study 1, our method outperformed alternative baseline models that use the Word2Vec and LIWC embeddings. These models achieved much lower correlations than SBERT in all three studies (though note that the performance of the Word2Vec method was very close to that of SBERT in Study 2B).

Furthermore, Figure 4 shows that the average out-of-sample predictive performance of our method was, for all questionnaires, on par with that of human judges incentivized to make accurate predictions. The average correlation across all experimental conditions for the 16PF was .40 ($t(598) = 37.02$, $p < .001$; 95%-CI .37 to 42) for our model, and .39 ($t(598) = 33.60$, $p < .001$; 95%-CI .37 to .42) for the human judges. For the RIASEC it was .38 ($t(598) = 18.29$, $p < .001$; 95%-CI .34 to .42) for our model, and .32 ($t(598) = 14.72$, $p < .001$; 95%-CI .28 to .37) for the human judges. For the HSQ it was .42 ($t(598) = 15.30$, $p < .001$; 95%-CI .37 to .48) for our model, and .42 ($t(598) = 15.86$, $p < .001$; 95%-CI .36 to .47) for the human judges. Our model performed slightly better than the human judges, albeit statistically non-significant for the 16PF (Welch's two-sample t-test, independent samples: $t(1152.59) = -0.08$, $p = .94$), RIASEC ($t(1060.20) = -1.95$, $p = .05$), and HSQ ($t(995.83) = -.18$, $p = .86$). Figures 5C-H show that most of the model and human performance is positive. The scatterplots show no outliers in model/human performance, except a single outlier for the 16PF, for which our model performs significantly worse than human judges (-1.0 vs .02). This outlier did not influence any of our results and model and human performance strongly to moderately correlated for all questionnaires (16PF: $r(58) = .70$, $p < .001$; RIASEC: $r(58) = .53$, $p < .001$; HSQ: $r(58) = .46$, $p < .001$) indicating that the prediction problems were similar for both model and humans.

Overall, the results of Studies 2A-C show that the success of our approach is not specific to the NEO-PI-R but instead generalizable to other common and less common questionnaires used in personality research.

Study 3

Although Study 1 and 2A-C provide evidence for the power of our approach, each of these studies uses a single personality questionnaire, composed of a small set of curated items.

As a more challenging test, Study 3 examined whether our method could achieve good performance on a large and unconstrained collection of items, spanning a diverse set of domains and constructs, as well as grammar and phrasing structures. For this, we collected a new dataset of personality ratings of over 3,500 different items taken from the International Personality Item Pool (IPIP; Goldberg et al., 2006).

Methods

161 participants (62.30% female; Mean age = 19.88) were recruited through the University's undergraduate subject pool. We selected the sample size for this study to ensure that we have data from at least 150 participants to evaluate our approach. Participants indicated at the beginning of the study if they wanted to answer one, two, or three blocks of approximately 300 items (each for a fixed amount of study credit). These items were sampled from a larger set of 3,563 personality items, all available items at the time of data collection in 2019, taken from the International Personality Item Pool (IPIP; Goldberg et al., 2006). The IPIP is a very large, broad collection of personality items across a multitude of different scales and constructs. It covers items from 36 distinct scales, involving 242 distinct constructs. For example, the IPIP includes constructs such as Tolerance, Adaptability, or Toughness with items such as "I believe in equality between all races", "I adapt easily to new situations", or "I remain calm under pressure". Note that the IPIP lists two item sets, one "total" list with 3,320 items and one "assigned only" list with items that have information about scales and constructs provided. We obtained 3,563 unique items after merging both lists. Also note that the number of scales and constructs referred to in this paper was taken from the items that have scale and construct information provided (see https://ipip.ori.org/ItemAssignmentTable.htm for more information). Each item involved

responses on a 5-point Likert scale. See Table 1 for a summary of dataset characteristics. Model predictions for Study 3 participants were obtained using the same methods as in Studies 2A-C.

Results and Discussion

Figures 2I and 2J show that the out-of-sample predictive performance of the SBERT method persisted for Study 3, with an average correlation of .37 ($t(160) = 44.24$, $p < .001$; 95%-CI .35 to .38). Additionally, our SBERT model achieved significant ($p < .05$) positive correlations for 99.4% of participants. Finally, as in Studies 1 and 2A-C, our method outperformed alternative baseline models that use the Word2Vec and LIWC embeddings. Overall, these results demonstrate the ability of our method to generalize across constructs and questionnaires, even beyond the standard questionnaire format. In other words, our approach can be used to make accurate predictions for thousands of different personality items using only a small set of participant ratings.

Study 4

Studies 1-3 have shown that our approach is able to predict participant responses to out-of-sample personality items, and moreover do so with a human-level of accuracy. This is likely because the SBERT semantic space captures the meanings of items in a manner that corresponds to the distribution of personality traits in the population. In other words, traits that correlate in the population are likely to have SBERT embeddings that are highly similar to each other. In Study 4, we rigorously tested this assumption by using item embeddings to predict the item's construct and its direction of loading on the construct. Note that this test is quite challenging (especially for the IPIP data), as the model has to learn how to classify items into a very large number of constructs.

Methods

The items for Study 4 were taken from Studies 1-3, and consisted of the NEO-PI-R, 16PF, RIASEC, HSQ and IPIP questionnaire (see Table 1 for a summary). Model training and prediction was done using a procedure that was similar to that in Studies 1-3. However, instead of learning to predict participant responses using item embeddings, our models learnt to predict the questionnaire construct that the item belonged to. This is a multinomial classification problem in which the number of categories corresponds to the number of constructs in the questionnaire. Thus, for example, for the NEO-PI-R questionnaire, our model attempted to predict whether a held-out item would fall into the Openness, Conscientiousness, Extraversion, Agreeableness, or Neuroticism categories. Importantly it attempted to do so using only the items' texts, and not human responses to the items.

We also considered a second model type that was trained to predict the direction (positive or negative) on which an item loaded onto a construct (e.g., "I feel comfortable around people" versus "I keep in the background" for the Extraversion construct in NEO-PI-R). We trained this model in a manner that was identical to the first model but used the items' directional loading as labels during classifier training. Note that we did not train this model on the RIASEC or IPIP data because these questionnaires do not have item direction codes (RIASEC has only positive directional loadings; the IPIP data does not provide directional loadings for most items).

To keep the model training procedure consistent across studies and tasks, we again identified the best performing machine learning technique and associated hyperparameters on the NEO-PI-R dataset for each embedding type (SBERT, Word2Vec, LIWC) and then applied the respective models and hyper-parameters on the remaining datasets. Note, since the labels in this task reflect a nominal scale (item construct), we only considered classification and not regression

algorithms from Study 1 (Ridge Classification, K-Nearest Neighbor Classification, Support Vector Classification). To get robust estimates of our model's out-of-sample performance we again applied cross-validation. Analogous to Studies 1-3, we then compared our predictive model against two alternative embeddings (Word2Vec and LIWC). To evaluate model performances, for both construct and direction prediction, we calculated the classification accuracy as the percentage of correct classifications across all predictions. Note that the prediction classes, personality constructs and directional loadings, are balanced for all questionnaires, except the IPIP which has a non-equal number of items for the different personality constructs.

Results

Semantic Similarity. Before building predictive models of construct classification, we first examined whether the embedding similarity of pairs of items from the same construct was (significantly) larger than that of items from different constructs. We measured the similarities between items by calculating the cosine similarity of their SBERT embeddings. We did this for all pairs of items in a questionnaire, then regressing pairwise item similarity scores onto a binary independent variable describing whether or not the items came from the same construct. We also included fixed effects for each of the constructs in these regressions.

The regression showed a significant positive coefficient for the in-construct variable for all datasets, except the HSQ ($b = 0.03$, $t(491) = 1.581$, $p = .114$). Details are presented in Table 2.

Table 2

Regression Results of Items' Semantic Similarity Over Construct Origin (Same vs Different)

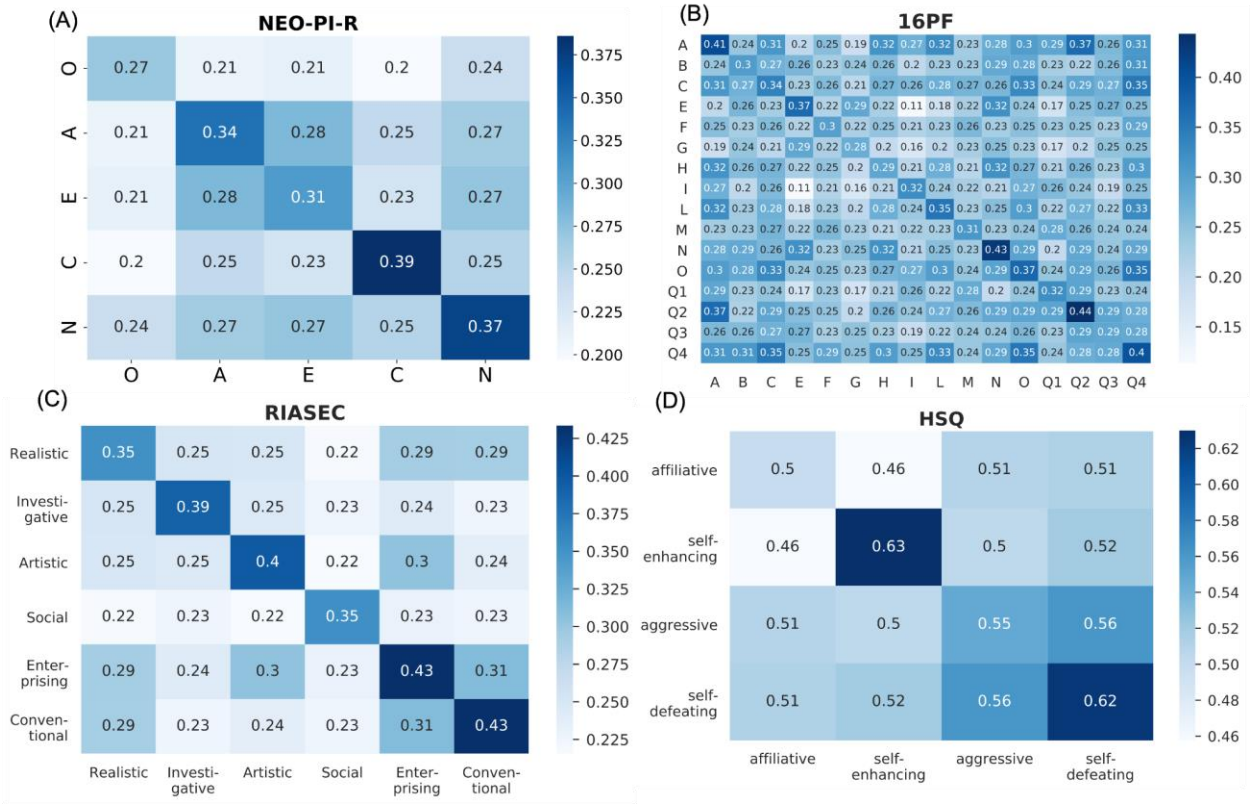
Dataset	*b*	*SE*	95%-CI	*t*	*p*
NEO-PI-R	0.069	0.006	[0.057, 0.081]	11.51	< .001
16PF	0.052	0.006	[0.04, 0.06]	8.35	< .001
RIASEC	0.130	0.010	[0.111, 0.149]	13.58	< .001
HSQ	0.030	0.019	[-0.007, 0.067]	1.58	.114

Note. We regressed item pairs' cosine similarity scores as a function of whether or not they are from the same construct, with fixed effects for each of the constructs. Results show that in-construct items significantly predicted higher cosine-similarity scores, except for the HSQ.

A visualization of the results is presented in Figure 6, for the NEO-PI-R data set (6A), the 16PF data set (6B), the RIASEC data set (6C), and the HSQ data set (6D). The IPIP dataset is omitted from Figure 6 since it has too many constructs (242 in total). The figure shows the average similarity scores for item pairs grouped by their constructs as a heatmap and indicates that, on average, items that load onto a given construct are closer to other items that load onto that same construct relative to items that load onto other constructs. In other words, personality items, across multiple distinct questionnaires, cohere together in stable constructs not only in human data but also in linguistic meaning.

Figure 6

Average Semantic Similarity Between Items in Each Pair of Constructs



Note. The similarity of each pair was calculated using the averaged cosine similarity between all items of one construct and all items of the other. Figure shows a general trend of higher in-construct vs between-construct similarity.

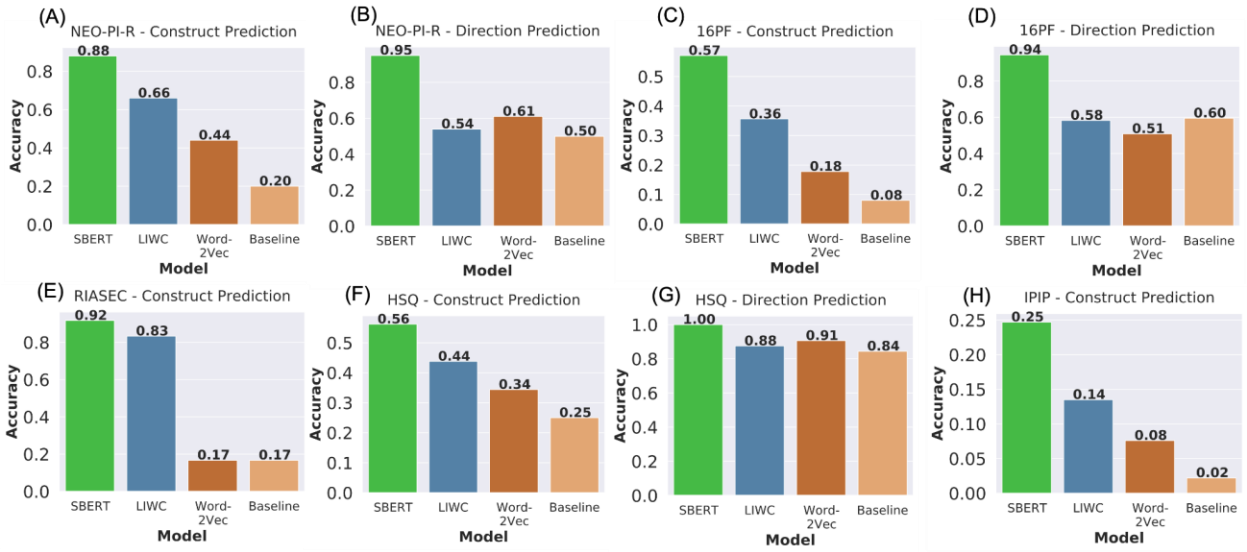
Figure 6 also shows some high cross-construct similarities. For example, items from the Agreeableness and Extraversion constructs of NEO-PI-R are highly similar to each other, as are items from the Aggressive and Self-defeating constructs of HSQ. However, it should be noted here that a) the same-construct similarity is still higher and b) our predictive models (in the subsequent analysis) use high-dimensional item representations and not a single similarity score to make predictions. As such, these models can use additional information, such as the exact positions or directions relative to other items in the embedding space, for predicting an item's construct and distinguishing it from other constructs that have semantically similar items.

Construct Prediction. We found the Ridge Classification model with a = 1,000 to be the best performing model on the NEO-PI-R and applied it on all other datasets as well. Figures 7A, 7C, 7E, 7F and 7H illustrate the results of the construct prediction task for each questionnaire. Here we see that our model achieved very high accuracy rates, outperforming both a random baseline (which would achieve an accuracy equaling the proportion of the most frequent construct) as well as the Word2Vec and LIWC baselines. This indicates that SBERT embeddings are indeed able to distinguish between same- and cross-construct items. Interestingly, our model achieved a 25% accuracy for the IPIP dataset, despite the very large number of constructs (a random model would achieve only 2% accuracy for this test). Figures 7B, 7D and 7G show similar results for the direction prediction task. For this task, we found the Ridge Classification model with a = 10 to be the best performing model on the NEO-PI-R and applied it on all other datasets as well. We achieved high accuracy rates, of up to 100%, across all questionnaires, outperforming every baseline.

We also analyzed the performance of our models for individual constructs. Our main model's prediction accuracy is high across all constructs with only few exceptions. For instance, the performance on four of the 16PF constructs (Emotional stability, social boldness, Apprehension, and Tension) is low (20%-50%) compared to the remaining constructs (70%-80%). This might be due to the high inter-correlation of these 16PF constructs, which were explicitly designed to be non-orthogonal (Saville & Blinkhorn, 1981). For the IPIP, our model performed well over a variety of constructs with an over representation of Interest related items in the top performing constructs and Expressivity related items in the bottom performing constructs.

Figure 7

Prediction Accuracy in the Construct and Direction Prediction Tasks for All Datasets



Note. The results show that SBERT item embeddings can accurately capture items' construct and direction information, and subsequently predict these variables with high accuracy. The baseline value corresponds to the accuracy achieved by a model that predicts constructs and directions randomly or as the majority class (for the IPIP).

It should be noted here, however, that the IPIP items stem from a multitude of scales and as such from a wide range of potentially very fine-grained constructs with strong conceptual overlap (see e.g., constructs, such as Tolerance, Compassion, Forgiveness, Mercy, etc.). Therefore, a lower performance of our model on some constructs does not necessarily indicate a systematic weakness in capturing these meanings. Indeed, a closer look into the misclassified items reveals that many were classified as closely related constructs. For instance, "I find it hard to forgive others" and "I try to forgive and forget" were misclassified as Forgiveness/Mercy instead of Compassion and "I express my affection physically" and "I have difficulty showing affection" were misclassified as Romantic disinterest instead of Positive expressivity. Similarly, some items that belong to overarching, higher-level constructs are misclassified because there are suitable lower-level constructs and facets that have a closer semantic match. For example, "I

break rules" and "I continue until everything is perfect" were misclassified as Norm-violation and Perfectionism instead of Conscientiousness. However, this is not captured by the accuracy metric, which ignores the closeness of the model's misclassifications to the true constructs. Yet, our model was still able to achieve accuracies several times above baselines indicating our approach's broad ability to capture psychological meanings in item texts which it can then leverage to make more accurate predictions.

Clustering. In our final analysis we attempted to interpret the latent structure of our SBERT item space by clustering embeddings for all 3,653 IPIP items into a small set of clusters. For this we used the K-Means clustering algorithm with $K = 5$, meaning that the items were grouped into five clusters. Note that this parameter is unrelated to the K in K-NN regression, which is a regularization parameter expressing how many responses to similar items (the K most similar ones) are used to estimate the response to another item. For K-Means clustering, K represents a top-down assumption on the number of clusters in the data and can be used for theorizing (i.e., by setting the number of clusters based on theoretical considerations). We chose $K = 5$ to show the feasibility of clustering large item sets into small sets of clusters and for ease of presentation. However, we are not bound to any specific clustering solution. The resulting clusters are presented in Table 3, which reports the most frequent traits measured by the items in the respective clusters. Here we can see that the clusters reflect reasonable psychological topics. For instance, cluster 1 refers to traits associated with attention problems (e.g., adhd and conscientiousness), with exemplary items containing "I felt like I was dreaming when I was awake", "I have been told I am not listening when others are speaking" and "I make careless mistakes". The remaining clusters 2, 3,4 and 5 refer to traits related to mental health (e.g., depression and anxiety), integrity (e.g., honesty and humility), leadership qualities (e.g.,

organization and leadership items), and sociality (e.g., sociability and sensation-seeking items) respectively. Table 4 provides an overview of exemplary items in each cluster.

Table 3

Summary of the Three Most Frequent Constructs Within Each Identified Cluster

Cluster	Trait 1	Trait 2	Trait 3
1	dissociation	adhd	conscientiousness
2	depression	anxiety	sentimentality
3	forgiveness/mercy	modesty/humility	honesty/integrity/authenticity
4	creativity/originality	organization	leadership
5	aesthetic appreciation/artistic interests	sociability	risk-taking/sensation-seeking/thrill-seeking

Note. We clustered all IPIP items based on their embeddings. Table shows the most frequent traits within each resulting cluster. Cluster 1 relates to attention problems. Cluster 2 relates to mental health. Cluster 3 relates to integrity. Cluster 4 relates to leadership qualities. Cluster 5 relates to sociality traits. The results show that we can use embeddings to cast unlabeled items, such as the ones in the IPIP inventory or newly created items, into reasonable constructs.

Although this demonstration is undoubtedly only a preliminary qualitative analysis, it shows that our approach can be used to meaningfully cluster large sets of items using only the item texts (see also Rosenbusch et al., 2020 for a similar test using word embeddings rather than sentence embeddings). Thus, this method can be used as a basis for future theoretical work that attempts to synthesize the very large set of items and behaviors that describe human personality.

Table 4

Exemplary Items Within Each Identified Cluster

Cluster Name	Item 1	Item 2	Item 3
1	I felt like I was dreaming when I was awake	I have been told I am not listening when others are speaking	I make careless mistakes
2	I am sad most of the time	I am generally a fearful person	I am easily moved to tears
3	I accept apologies easily	I do not call attention to myself	I am true to my own values
4	I come up with new ways to do things	I supervise the work of others	I am good at helping people work well together
5	I can become tearful thinking of the goodness of others	I am interested in people	I prefer friends who are excitingly unpredictable

Note. We clustered all IPIP items based on the similarity of their vector representations. Table shows exemplary items for each identified cluster. All items presented belong to the clusters' most frequent traits as shown in Table 3. The table provides an overview of the specific measures clustered together based on the semantic relations of the item texts.

Discussion

While Studies 1-3 evaluated the utility of SBERT embeddings for predicting participant ratings, Study 4 attempted to use SBERT item embeddings to classify items into constructs. Using the questionnaires from Studies 1-3, we found that items from the same construct were significantly more similar to each other in the SBERT embedding space than items from different constructs. Additionally, we found that our machine learning models could accurately predict both the construct and the loading (positive or negative) of items onto the construct in a questionnaire. Again, models based on SBERT embeddings outperformed those based on Word2Vec or LIWC, showing that SBERT provides a superior representation of item meaning.

Finally, we provided a simple demonstration of the utility of our SBERT embeddings for construct classification, by clustering all 3,653 IPIP items into a small number of categories based on their positions in the SBERT embedding space. Together, the tests conducted in Study 4 show that item embeddings are able to capture the psychological content of questionnaire items using only their text.

General Discussion

The lexical hypothesis proposes that personality traits that are important to a group are encoded in words of its everyday language (Goldberg, 1990; John et al., 1988; Klages, 1929). This hypothesis has provided a theoretical basis for the use of linguistic descriptors in personality research, and in this way, has guided and constrained personality research for decades. Expanding on the lexical hypothesis, we have used a recent deep language model, SBERT (Reimers et al., 2019; see also Devlin et al., 2019), to obtain representations for items in personality questionnaires. SBERT is trained on large amounts of natural language data and can accurately represent the meaning of sentences in high-dimensional embeddings that are sensitive to word order and syntax. We have used SBERT embeddings, combined with various machine learning models, to predict participant responses to out-of-sample personality items from several existing questionnaire datasets: NEO-PI-R (Stillwell & Kosinski. 2012) in Study 1, 16PF (Cattell & Mead, 2008) in Study 2A, RIASEC (Liao et al., 2008) in Study 2B, and HSQ (Martin et al., 2003) in Study 2C. Across these studies we have found that the SBERT approach achieves high accuracy rates, greatly outperforming accuracy rates obtained from baseline models that represent questionnaire items using either Word2Vec word embeddings (Mikolov et al., 2013) or Linguistic Inquiry and Word Count (LIWC) dimensions (Pennebaker et al., 2015). Word2Vec can capture word meaning but does not take into account the structure and word order in

personality items. Likewise, LIWC uses word frequency statistics to categorize items on several psychological variables but does not encode additional nuances in the meanings of the items. The superior performance of SBERT relative to Word2Vec and LIWC highlights the importance of sentence embeddings for capturing the meanings of personality items, and the overall value of recent advances in deep learning and natural language processing for the prediction of human personality.

Studies 1 and 2A-C have used existing personality questionnaires to measure model accuracy. To test whether our approach can also apply to a larger unconstrained set of personality items, we collected new data in Study 3, in which we offered over 3,500 personality items from the International Personality Item Pool (IPIP; Goldberg et al., 2006) to human participants. As with our previous studies, we found that our approach successfully extrapolates personality ratings and predicts people's responses even when test items involve diverse domains and different types of constructs. This illustrates the broad applicability of our method for personality prediction.

In Study 1 and Study 2A-C we also compared the performance of our approach to incentivized human judges that were given an identical prediction task. We found that our approach achieved similar performance to human judges, indicating that the accuracy levels documented in this paper match the accuracy levels that can be obtained by humans. Additionally, there was a high correlation between human accuracy and model accuracy in all studies, indicating that target participants that were easy or difficult to predict by human judges were also easy or difficult to predict by the model. This provides strong evidence that our model is able to represent the meanings of personality items in a human-like manner.

In Study 4 we directly tested whether the SBERT assigned similar representations to closely related personality traits. We found that this was indeed the case, with items belonging to the same construct being given similar embeddings. For this reason, our approach was able to predict the construct that an item belonged to and could even describe how an item loaded onto its construct. Again, SBERT was much better at construct prediction than Word2Vec and LIWC. SBERT was also able to generate interpretable clusters of IPIP items in a bottom-up manner, indicating that it can be used for scale construction and construct delivery.

The success of our approach has implications for the representation and measurement of personality. Prior research has represented a person's personality using a collection of linguistic descriptors. For example, a person may be described as being high in extraversion, where extraversion is specified as a set of sentences or trait words. We show that these linguistic descriptors can themselves be assigned quantitative representations. Thus, we are able to represent an individual using a collection of points (corresponding to the personality items they rate highly) in a high-dimensional space. Thus, even though the embeddings used in our analysis are the representations of the language models, since these language models represent the meanings of sentences, and sentences are (in accordance with the lexical hypothesis) used to represent personality, the embeddings by proxy also become representations of personality. The advantage of these quantitative representations is that they also inherently express the relationship between items and constructs (e.g., via distance metrics in the embedding space), as shown in the construct prediction and the clustering tasks.

Testing whether our approach extends to other psychological variables is an important topic for future work. For example, life satisfaction is a variable associated with several personality traits (Schimmack et al. 2002, Schimmack et al., 2004, Steel et al., 2008). The study

of the interplay between life satisfaction and personality currently involves regressing participants' ratings of life satisfaction on a small handful of personality dimensions (e.g., those from the five-factor model). Instead, using our approach, we could use a participant's position in our language space as a predictor, by treating each individual as the sum of the embeddings of the items that they rate highly. The high dimensionality of this representation could lead to more accurate predictions and shed light on the specific personality items (i.e., regions of embedding space) that correlate with life satisfaction (see Nimon et al., 2016 for a related analysis). Of course, similar analyses could be attempted with other important psychological variables, as well as socio-economic, neural, and even genetic variables (Ayoub et al., 2018; Kennis et al., 2013; McGue et al., 1993; Sugiura et al., 2000; Vukasović et al., 2015). Attempting such an analysis is a fruitful topic for future work.

Our approach is complementary to recent work that attempts to model personality using high-dimensional feature vectors for individuals obtained from their social media activity (Bleidorn & Hopwood, 2019; Kosinski et al., 2013; Park et al., 2015; Stachl et al., 2021). By featuring individuals, these researchers can successfully predict responses to personality surveys for out-of-sample individuals. For example, it is now possible to determine, based only on a person's social media activity, whether a person will be high or low on extraversion. While social media analyses focus on predicting out-of-sample individuals, we show that it is also possible to predict out-of-sample personality items. By combining the two approaches it may even be possible to predict responses of out-of-sample individuals to out-of-sample items. If successful, this would help personality researchers characterize arbitrary individuals on arbitrary traits, and thus open the door to several new types of quantitative behavioral analyses (see Vu et al. (2020) for recent work demonstrating the feasibility of this analysis). However, these

approaches need to be carefully balanced with ethical considerations. Researchers applying our methods must ensure strict observance of informed consent, opt-out of data usage and especially data anonymization to prevent out-of-sample participant predictions from being applied on unwilling participants (e.g., by collecting their online data) and out-of-sample item predictions from being applied on sensitive items that participants refuse to answer (e.g., questions on sexual orientation).

Quantifying personality items, as we have done in this paper, also has value beyond prediction. For example, we have shown that our methods can be used to infer high-level constructs associated with unlabeled personality items, and even cluster large sets of items into a smaller set of constructs. Importantly, this method does not require participant data, making it easily scalable to tens of thousands of personality items. This method has particular value for survey design. For example, it would be possible to optimize the set of personality items used in a given survey, by selecting personality items from distinct regions of language space, thereby maximizing the information acquired in a survey (Arnulf et al., 2014; Evans et al., 2022; Garcia et al., 2020; Rosenbusch et al., 2020). Similar optimal experimental design methods have been proposed in other areas of psychology (e.g., Myung & Pitt, 2009), and can be extended to personality research, as we are now able to quantitatively represent the stimuli used in personality experiments (see Yarkoni, 2010).

It should be noted here that our approach relies on data exclusively collected from western samples. The over-reliance on so-called WEIRD (western, educated, industrialized, rich, democratic) samples for psychology research has been criticized and previous research found that many fundamental cognitive and affective processes differ across populations (see Henrich et al., 2010). Similarly, the language models used to create quantitative item representations are

based on the language use of mostly WEIRD populations in English. Different populations and languages might have different linguistic conceptualizations of personality that, following the lexical hypothesis, should map better on these populations' personalities (see Laajaj et al. (2019) for difficulty accessing personality traits using WEIRD-based questionnaires). As such, our models could be WEIRD-skewed and mask potential group-specific differences in personality traits. This gives reason for caution against an unreflective use of our models and should motivate researchers to consider language models trained on non-WEIRD corpora. For instance, ParsBERT (Farahani et al., 2021) and ARBERT/MARBERT (Abdul-Mageed et al., 2021) are monolingual language models trained on Persian and Arabic corpora and subsequently outperform multilingual language models usually used for these languages (e.g., mBERT or XLM-R, Conneau et al., 2020). However, the cultural bias in the language models is also an opportunity to quantitatively study group differences using this framework. For instance, to what extent do item representations based on different languages or language use of different populations differ? Do cultural perceptions and connotations of personality in language map onto measurements of personality? Another potentially fruitful line of research could incorporate current debiasing efforts of language models (e.g., Barikeri et al., 2021; Lauscher et al., 2020; Liang et al., 2020) into this approach. Reducing cultural biases in language models could help increase model performance and generalizability to non-WEIRD populations.

Finally, the applicability of our approach is not limited to personality research. For example, it would be possible to apply an identical research pipeline to predict emotion and well-being, using the thousands of items across hundreds of survey questionnaires and well-being dimensions, that have been proposed in prior research (Linton et al., 2016). A similar approach could be applied to the prediction of risk preferences, organizational attitudes, and health

judgments (Aka & Bhatia, 2022; Bhatia, 2019; Bhatia et al., 2022; Gandhi et al., 2022; Singh et al., 2022). More generally, survey-based research is a central component of many areas in the behavioral and cognitive sciences, including health, clinical, consumer, and managerial psychology. We show that it is possible to accurately quantify the language in surveys with deep networks, providing researchers with mathematical representations for verbal constructs and items. We look forward to future work that extends our approach to alternate domains in the behavioral sciences, thereby facilitating not only more accurate prediction, but also more rigorous scientific theorizing.

References

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7088–7105. doi:10.18653/v1/2021.acl-long.551

Aka, A., & Bhatia, S. (2022). Machine Learning Models for Predicting, Understanding, and Influencing Health Perception. Journal of the Association for Consumer Research, 7(2), 142-153.

Allport GW, Odbert HG. Trait names: a psycholexial study. Psychological Monographs. 1936;47:1. doi: 10.1037/h0093399

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3.

Armstrong, P. I., Day, S. X., McVay, J. P., & Rounds, J. (2008). Holland's RIASEC model as an integrative framework for individual differences. Journal of Counseling Psychology, 55(1), 1.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. PloS one, 9(9), e106361.

Ayoub, M., Gosling, S. D., Potter, J., Shanahan, M., & Roberts, B. W. (2018). The relations between parental socioeconomic status, personality, and life outcomes. Social Psychological and Personality Science, 9(3), 338-352.

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1941–1955. doi:10.18653/v1/2021.acl-long.151

Bhatia, S. (2019). Predicting risk perception: New insights from data science. Management Science, 65, 3800-3823.

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (in press). Predicting leadership perception with large-scale natural language data. Leadership Quarterly.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. Personality and Social Psychology Review, 23(2), 190-203.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 632–642. doi:10.18653/v1/D15-1075

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems.

Cattell, H. E., & Mead, A. D. (2008). The sixteen personality factor questionnaire (16PF). The SAGE handbook of personality theory and assessment, 2, 135-159.

Cattell RB. The description of personality I. Foundations of trait measurement. Psychological Review. 1943;50:559–594. doi: 10.1037/h0057276.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 1–14. doi:10.18653/v1/S17-2001

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. doi:10.18653/v1/2020.acl-main.747

Conneau, A., & Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). https://aclanthology.org/L18-1269

Costa, P. T., & McCrae, R. R. (1992). Neo personality inventory-revised (NEO PI-R). Odessa, FL: Psychological Assessment Resources.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. doi:10.18653/v1/N19-1423

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual review of psychology, 41(1), 417-440.

Evans, A. M., Rosenbusch, H., & Zeelenberg, M. (2022). Using Semantic Similarity to Understand the Psychological Constructs Related to Prosociality. Current Opinion in Psychology.

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). Parsbert: Transformer-based model for persian language understanding. Neural Processing Letters, 53(6), 3831-3847.

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. Psychological Science, 33(4), 579-594.

Galton, F. (1884). Measurement of character. Fortnightly Review, 36, 179-185. Fortnightly Review. 1884.

Garcia, D., Rosenberg, P., Nima, A. A., Granjard, A., Cloninger, K. M., & Sikström, S. (2020). Validation of two short personality inventories using self-descriptions in natural language and quantitative semantics test theory. Frontiers in psychology, 11, 16.

Golbeck, J. A. (2016). Predicting personality from social media text. AIS Transactions on Replication Research, 2(1), 2.

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. Journal of personality and social psychology, 59(6), 1216.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. Journal of Research in personality, 40(1), 84-96.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. European journal of Personality, 2(3), 171-203.

Kennis, M., Rademaker, A. R., & Geuze, E. (2013). Neural correlates of personality: an integrative review. Neuroscience & Biobehavioral Reviews, 37(1), 73-95.

Klages, L., & Johnson, W. H. (1929). The science of character. Mind, 38(152).

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences, 110(15), 5802-5805.

Laajaj, R., Macours, K., Pinzon Hernandez, D. A., Arias, O., Gosling, S. D., Potter, J., ... & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. Science advances, 5(7), eaaw5226.

Lauscher, A., Glavaš, G., Ponzetto, S. P., & Vulić, I. (2020). A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. Proceedings of the AAAI

Conference on Artificial Intelligence, 34(05), 8131-8138. https://doi.org/10.1609/aaai.v34i05.6325

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871–7880. doi:10.18653/v1/2020.acl-main.703

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., & Morency, L.-P. (2020). Towards Debiasing Sentence Representations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5502–5515. doi:10.18653/v1/2020.acl-main.488

Liao, H. Y., Armstrong, P. I., & Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. Journal of Vocational Behavior, 73(1), 159-183.

Linton, M. J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: exploring

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. (2016, March). Analyzing personality through social media profile picture choice. In Tenth international AAAI conference on web and social media.

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.

dimensions of well-being and developments over time. BMJ open, 6(7), e010641.

Lopez, S. J., & Snyder, C. R. (2003). Positive psychological assessment: A handbook of models and measures (pp. xvii-495). American Psychological Association.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. Journal of research in personality, 37(1), 48-75.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. Journal of personality, 60(2), 175-215.

McGue, M., Bacon, S., & Lykken, D. T. (1993). Personality stability and change in early adulthood: A behavioral genetic analysis. Developmental psychology, 29(1), 96.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. arXiv:1301.3781v1

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. Psychological review, 116(3), 499.

Nimon, K., Shuck, B., & Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence?. Journal of Happiness Studies, 17(3), 1149-1171.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6), 934.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. doi:10.18653/v1/D19-1410

Richie, R. & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. Cognitive Science, 45(8), 13030.

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. Psychological methods, 25(3), 380.

Rossier, J., Meyer de Stadelhofen, F., & Berthoud, S. (2004). The Hierarchical Structures of the NEO PI-R and the 16PF5. European Journal of Psychological Assessment, 20(1), 27.

Saville, P., & Blinkhorn, S. (1981). Reliability, homogeneity and the construct validity of Cattell's 16PF. Personality and individual differences, 2(4), 325-333.

Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: integrating process models of life satisfaction. Journal of personality and social psychology, 82(4), 582.

Schimmack, U., Oishi, S., Furr, R. M., & Funder, D. C. (2004). Personality and life satisfaction: A facet-level analysis. Personality and social psychology bulletin, 30(8), 1062-1075.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017, September). Dlatk: Differential language analysis toolkit. In Proceedings of the 2017

conference on empirical methods in natural language processing: System demonstrations (pp. 55-60).

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

Singh, M., Richie, R. & Bhatia, S. (2022). Representing and predicting everyday behavior. Computational Brain & Behavior, 5, 1-21.

Stachl, C., Boyd, R. L., Horstmann, K. T., Khambatta, P., Matz, S., & Harari, G. M. (2021). Computational Personality Assessment-An Overview and Perspective.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. Psychological bulletin, 134(1), 138.

Stillwell, D. J., & Kosinski, M. (2012). myPersonality project: Example of successful utilization of online social networks for large-scale social research. Proceedings of the ACM Workshop on Mobile Systems for Computational Social Science (MobiSys)

Sugiura, M., Kawashima, R., Nakagawa, M., Okada, K., Sato, T., Goto, R., ... & Fukuda, H. (2000). Correlation between human personality and neural activity in cerebral cortex. Neuroimage, 11(5), 541-546.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Vu, H., Abdurahman, S., Bhatia, S., & Ungar, L. (2020, November). Predicting responses to psychological questionnaires from participants' social media posts and question text

embeddings. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 1512-1524).

Vukasović, T., & Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. Psychological bulletin, 141(4), 769.

Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. Journal of research in personality, 44(2), 180-198.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100-1122.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036-1040.

Appendix

Table A1

Prediction Accuracy on the NEO-PI-R Dataset for Models Using SBERT Embeddings

Number	Model	Hyperparameter	Correlation	95%-CI
1	Ridge Classification	a = 1	.34	[.33, .34]
2		a = 10	.34	[.33, .35]
3		a = 100	.37	[.36, .37]
4		a = 1,000	.41	[.41, .42]
5		a = 10,000	.35	[.34, .35]
6	Ridge Regression	a = 1	.38	[.37, .38]
7		a = 10	.38	[.37, .39]
8		a = 100	.41	[.40, .41]
9		a = 1,000	.45	[.44, .45]
10		a = 10,000	.27	[.26, .27]
11	K-NN Classification	K = 1	.37	[.37, .38]
12		K = 5	.41	[.40, .41]
13		K = 10	.38	[.37, .38]
14		K = 15	.37	[.37, .38]
15		K = 50	.20	[.19, .21]
16	**K-NN Regression**	K = 1	.37	[.37, .38]
17		**K = 5**	**.45**	**[.45, .46]**
18		K = 10	.42	[.41, .42]
19		K = 15	.40	[.40, .41]
20		K = 50	.14	[.13, .14]
21	SVC	C = 1	.41	[.40, .41]

22		C = 10	.43	[.43, .44]
23		C = 100	.43	[.43, .44]
24		C = 1,000	.43	[.43, .44]
25		C = 10,000	.43	[.43, .44]

Note. This table compares the accuracy of different machine learning techniques over a range of hyperparameters. Accuracy is measured in terms of average correlation across N = 2,749 study participants. The best performing model (K-NN regression, K =5), which has been bolded, was applied to other questionnaires: 16PF, RIASEC, HSQ, and IPIP in Studies 2 and 3. All correlations are significantly different to zero at $p < .001$, with Bonferroni corrected significance level of .002.