That Takes the BISCUIT

Predictive Accuracy and Parsimony of Four Statistical Learning Techniques in Personality Data, With Data Missingness Conditions

Lorien G. Elleman[1], Sarah K. McDougald[1], David M. Condon[2], and William Revelle[1]

[1]Department of Psychology, Northwestern University, Evanston, IL, USA
[2]Department of Psychology, University of Oregon, Eugene, OR, USA

Abstract: The predictive accuracy of personality-criterion regression models may be improved with statistical learning (SL) techniques. This study introduced a novel SL technique, BISCUIT (Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent). The predictive accuracy and parsimony of BISCUIT were compared with three established SL techniques (the lasso, elastic net, and random forest) and regression using two sets of scales, for five criteria, across five levels of data missingness. BISCUIT's predictive accuracy was competitive with other SL techniques at higher levels of data missingness. BISCUIT most frequently produced the most parsimonious SL model. In terms of predictive accuracy, the elastic net and lasso dominated other techniques in the complete data condition and in conditions with up to 50% data missingness. Regression using 27 narrow traits was an intermediate choice for predictive accuracy. For most criteria and levels of data missingness, regression using the Big Five had the worst predictive accuracy. Overall, loss in predictive accuracy due to data missingness was modest, even at 90% data missingness. Findings suggest that personality researchers should consider incorporating planned data missingness and SL techniques into their designs and analyses.

Keywords: statistical learning, machine learning, personality, nuances, Big Five

Research over the last decade has indicated that personality items (often called "nuances"; McCrae, 2015) are both reliable and valid measures of personality. There is cross-rater agreement associated with the specific variance of nuances (Mõttus, McCrae, Allik, & Realo, 2014) and nuances have rank-order stability over time, and are heritable (Mõttus, Kandler, Bleidorn, Riemann, & McCrae 2017; Mõttus et al., 2019). Additionally, personality-criterion models that utilize nuances tend to be more predictive than those that employ broad domains (e.g., the Big Five; Goldberg, 1990) or narrower facets (Mõttus, Bates, Condon, Mroczek, & Revelle, 2017; Mõttus et al., 2015; Seeboth & Mõttus, 2018).

Item-level analysis requires a number of multiple comparisons that is an order of magnitude greater than broad personality domains or narrower facets. Traditional methods of analysis, such as regression, can overfit the data or find few stable results after statistical adjustments. Recently, several researchers have suggested using statistical learning (SL) techniques[1] to study nuances (Chapman, Weiss, & Duberstein, 2016) and improve the prediction of outcomes

in personality psychology (Yarkoni & Westfall, 2017). Compared to traditional statistical methods, many SL techniques are more complex and better suited to the study of nuances because they have been designed to reduce overfitting. Usually, the accuracy of an SL model is measured by the prediction of a hold-out sample (the "test sample") that has been kept separate from the sample upon which the model was built (the "training sample"). For an overview of statistical learning, see James, Witten, Hastie, and Tibshirani (2017). For short overviews, see Chapman et al. (2016) and Yarkoni and Westfall (2017).

To improve prediction of the test sample, an SL technique may augment a basic statistical method, such as regression, in several ways. For instance, an SL technique may implement "regularization" to shrink the coefficients of a model to reduce overfitting (e.g., ridge regression; Hoerl & Kennard, 1970). Some SL techniques use "variable selection" to retain the most important variables for the final model (e.g., the lasso; Tibshirani, 1994). SL techniques may test many different models via "resampling," an

---

[1] Specifically, supervised learning. Models generated by supervised learning techniques are "supervised" by the criterion variable they predict. Unsupervised learning techniques describe patterns in data without the use of a criterion.

iterative sampling procedure: each new model is developed iterative on randomly selected sub-samples of the training data and may be cross-validated using hold-out portions of the training data (for a review of using cross-validation for model selection, see Arlot & Celisse, 2010). Resampling procedures may be used to aggregate the different models into a final model, to efstimate the error of the model estimates, and/or to optimize model hyperparameters (or "tuning parameters"). A tuning parameter differs from a typical model parameter in that the researcher preselects a series of tuning parameter coefficients. Each tuning parameter coefficient is input into a new model or series of models. Hyperparameters are tuned (i.e., an optimal value is found for each) by selecting the model or aggregated model with the lowest cross-validated error. For example, the lasso's regularization hyperparameter must be tuned in order to determine the optimal degree of regularization for a particular criterion (Tibshirani, 1994).

Applying certain SL techniques to personality psychology may result in final models that are substantially more complex, and perhaps more difficult to interpret, than traditional personality models. For example, in applying an SL technique to personality data, Seeboth and Mõttus (2018) took an approach that was similar to a genome-wide association study (GWAS; Hirschhorn & Daly, 2005), such that personality-criterion associations were considered to be "driven by a large number of specific personality characteristics" (p. 188) and nuance-criterion relationships were summarized by the variance explained by using an unspecified number of items. Even if nuances predict a criterion better than facets or domains, certain SL methods, such as a "persome"-wide association study (Mõttus et al., 2017), may output a model with as many or nearly as many predictors as there are items in the pool. While predictive accuracy and parsimony differ for each SL approach, very little, if any, research in personality psychology has been performed to compare the predictive accuracy and parsimony of SL techniques.

The Four Statistical Learning Techniques to Be Compared

BISCUIT

The Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent (BISCUIT; Revelle, 2019), is a correlation-based SL technique that grew out of the practical need for generating parsimonious models to describe nuance-level relationships in Massively Missing Completely At Random (MMCAR) data (Revelle, Wilt, & Rosenthal, 2010; Revelle et al., 2016). In MMCAR data, each participant is given a random sample of items; the raw data are mostly (i.e., massively) missing, but this missingness has been completely randomized. Individual scales may be over- or undersampled.

Similar to the "criterion-keyed scale construction" of Chapman et al. (2016) and reminiscent of the procedures used in the development of the MMPI (Hathaway & McKinley, 1942), BISCUIT utilizes variable selection to retain the items that most strongly correlate with a criterion (i.e., the best items). Item-level correlations in BISCUIT are calculated solely from pairwise administrations of items. Thus, unlike other SL techniques in this study, BISCUIT may be run on MMCAR data structures without the need for imputation. BISCUIT uses a resampling procedure to determine a cross-validated list of the best items based upon the average correlation; either bootstrap aggregation ("bagging") or k-fold cross-validation may be utilized (for a description of bagging, see Breiman, 1996; for k-fold cross-validation, see Chapman et al., 2016, p. 607). The cross-validated best items are combined into a scale for the criterion, which is the final model for BISCUIT. In BISCUIT's empirically constructed scale (and typical personality scales), all best items are weighted the same (i.e., unit-weighted).[2]

Compared to an optimally weighted regression model, a unit-weighted model tends to fit the initial dataset about as well (e.g., Dawes, 1979; Wilks, 1938), and often has improved predictive accuracy in new datasets (Wainer, 1976; Waller, 2008); optimal weights are optimal only for the initial dataset, and overfitted in others. Although there is only one set of optimal weights for a least-squares regression model, there are an infinite number of alternative sets of weights for a more robust, non-least-squares solution (Waller, 2008). BISCUIT employs unit-weighting as a simple alternative to least-squares regression for the same reason that regression-based statistical learning techniques implement regularization: to improve upon the predictive accuracy of an overfitted regression model by systematically modifying the model's coefficients. Lastly, BISCUIT's unit-weighted models and output are like oven windows through which one can view a biscuit baking; BISCUIT outputs a list of items that most highly correlate with a criterion, their correlations with the criterion, and the content of each item. BISCUIT's tuning parameter is the number of best items to select for a model.

---

[2] Reviewers were concerned that BISCUIT's performance would improve by weighting variables instead of unit-weighting them. An option to weight variables (equal to their zero-order correlations) has been added to the BISCUIT algorithm. Comparative analysis indicated that BISCWIT (Weighted, instead of Unweighted) performed sometimes better than BISCUIT, sometimes worse, and on average about the same (see Table 13 in ESM 1). A reviewer commented that BISCWIT's performance could improve if its coefficients were estimated by multiple regression instead of zero-order correlation. We agree that exploring this modification in a future study would be worthwhile.

To provide clarity around the BISCUIT algorithm, the following is a step-by-step procedure for it:

(1) At least two options are selected: (1a) the range of *N* best items to be retained and (1b) whether the analysis should use bagging or *k*-fold cross-validation (this example will assume *k*-fold).

(2) For a given criterion, for each of *k* splits: (2a) A criterion-by-item correlation matrix is calculated, based on the pairwise administrations of the raw data in the training subsample. (2b) The *N* items that have the largest correlations with the criterion are retained and formed into a unit-weighted scale. Both item-level and scale-level correlations are recorded. (2c) The holdout subsample may be used to determine the cross-validated correlation of the unit-weighted scale with the criterion.

(3) The steps in 2 are repeated *k* times.

(4) Average correlations across the *k* splits are found.

(5) A final set of *N* items are retained, based on the number of items that were best cross-validated across the *k* splits.

(6) The BISCUIT model is output as a scale, listing each item and whether it is negatively or positively associated with the criterion.

Lasso

The Least Absolute Shrinkage and Selection Operator (lasso; Tibshirani, 1994) is a regression-based SL technique that was created to be an improvement over traditional regression and ridge regression (Hoerl & Kennard, 1970). The lasso and ridge regression are similar in that each uses a regularization penalty that is based on a tuning parameter and the magnitude of each regression coefficient. However, ridge regression's penalty (l_2) uses the square of each coefficient, while the lasso's penalty (l_1) uses the absolute value of each coefficient (see Equations 1 and 2 in Electronic Supplementary Material, ESM 1). The lasso's penalty, unlike ridge regression's penalty, allows regression coefficients to shrink to values of zero. After regularization, variables with zero-value coefficients are discarded, effectively giving the lasso a variable selection feature. The lasso's tuning parameter λ determines the magnitude of coefficient shrinkage.

Elastic Net

The elastic net is a regression-based SL technique that is framed as an improvement over the lasso (Zou & Hastie, 2005). The elastic net incorporates ridge regression and the lasso into one algorithm; the lasso is a special case of the elastic net when the λ_2 tuning parameter of the elastic net is set to zero, and ridge regression is a special case of the elastic net when λ_2 is set to 1 (Zou & Hastie, 2005).

Two typical tuning parameters of the elastic net are: (a) λ, which determines the magnitude of coefficient shrinkage; and (b) λ_2, which determines the extent to which groups of highly correlated variables will be retained.

Random Forest

The random forest (Breiman, 2001) is an SL technique based upon decision trees. A decision tree iteratively partitions a dataset, one variable at a time, into two groups such that differences in the groups maximally predict a criterion. Essentially, the random forest combines the bagging resampling procedure with the random decision forest (Ho, 1995). In the random decision forest, a final model is built from an aggregation of multiple trees; in each tree, a random subset of predicting variables is selected for each branch. The random forest combines bagging and the random decision forest by aggregating bootstrapped decision tree models, where each model includes a subsample of predicting variables. The purpose of bagging and the random decision forest is similar: to aggregate models based upon samples from the available data in order to reduce overfitting. There are inconsistencies in the literature regarding what, if any, tuning parameters should be used for the random forest (Probst & Boulesteix, 2018; Tang, Garreau, & von Luxburg, 2018).

Aims of the Study

The primary aim of this study was (a) using personality data, to compare the models of four SL techniques in terms of their predictive accuracy. Because of our particular interest in BISCUIT, and because BISCUIT was built to perform well with MMCAR data, we also evaluated, (b) in terms of predictive accuracy, whether BISCUIT models gained an advantage over other SL models as the rate of data missingness was artificially increased in the sample. Finally, we determined (c) the extent to which BISCUIT tended to provide more parsimonious models than other SL techniques, which was quantified by the number of personality items used in a model.

Methods

Sample

Participant data were collected at https://sapa-project.org, an international online personality assessment. The SAPA (Synthetic Aperture Personality Assessment) Project is an ongoing research project where each participant is given a small random sample of a large item pool (over 6,000 items), resulting in an MMCAR data structure. An initial

sample of 497,048 participants (64% female; Mdn_{age} = 26 years; from 228 countries; 39% from the US) was collected from February 7, 2017 to November 12, 2018. In order to run out-of-the-box algorithms for the lasso, elastic net, and random forest, the data were limited to complete cases for the selected personality items and criteria (see below). Requiring complete data reduced the sample to 78,828 participants. In the final sample, participants were from 200 countries (57% from the US), 65% were female, and the median age was 33 years (min = 14, max = 90). Descriptive information concerning the initial and final samples are available in Table 1 in ESM 1.

Measures

All measures were self-reported. Personality was measured with the 135-item SPI-27 (SAPA Personality Inventory; Condon, 2018), a personality inventory that may be scored as 27 traits (five items per trait) or as the Big Five domains (70 total items; 14 items per trait). Each personality item was answered on a 6-point Likert-like scale. There were five criteria: Body Mass Index (BMI), smoking frequency, sleep quality, general health, and educational achievement. These specific criteria were selected for their breadth. Demographic measures included ethnicity (if the participant was from the US), age, sex, and country of residence.

Procedure

All steps in the procedure and analyses were performed with the statistical programing language and environment R (R Core Team, 2019) in the integrated development environment RStudio (RStudio Team, 2019). There were three primary steps to preparing the data for analysis (Figure 1): (a) split the final sample into the test and training samples; (b) create new test and training sample datasets by imposing increasing levels of missingness; and (c) for each dataset with missing data, create new datasets in which the missing data were imputed. More details of each step are described below:

(a) The final sample was randomly split into the training sample (75% of participants) and test sample (the remaining 25%). Having the training sample be larger than the test sample gives training models greater power and is typical (e.g., Breiman, 1996; Chapman et al., 2016; Seeboth & Mõttus, 2018).

(b) Because BISCUIT was designed to analyze MMCAR data, it was necessary to test whether missingness in personality data would give an advantage to BISCUIT's predictive accuracy over the models of other techniques. To do this, four new datasets were created (for each of the training and test samples), where each
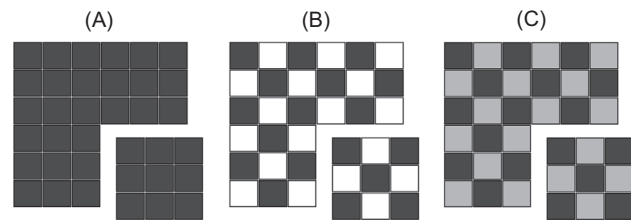

Figure 1. A visual representation of the three steps in which the sample data were prepared for analyses. (A) The final sample (complete data) was randomly split into the training sample (75% of the sample) and the test sample (25% of the sample). (B) For both the training and test samples, new datasets were created in which random missingness was imposed in the personality data. This representation only shows a dataset in which 50% missingness was imposed, but this procedure was also performed for 25%, 75%, and 90% missingness. (C) For each dataset with missing personality data, a new dataset was created in which the missing data were imputed. For levels of missingness in which multiple imputation was used, 20 datasets were created for each dataset with missing data.

new dataset imposed increasing levels of random missingness in the personality data (25%, 50%, 75%, and 90% missingness; see Table 2 in ESM 1 for pairwise administrations at each level of data missingness).

(c) BISCUIT's algorithm can converge on datasets with missing data, but other out-of-the-box SL techniques cannot. Therefore, new datasets were created that imputed the imposed missing data (using the "MIPCA" and "imputePCA" functions of the R package "missMDA"; Josse & Husson, 2012, 2016). For datasets with 25%, 50%, and 75% data missingness, imputation was performed with multiple imputation using Bayesian principal components analysis (BayesMIPCA; Audigier, Husson, & Josse, 2014). This imputation method performs favorably compared to other methods (Schmitt, Mandel, & Guedj, 2015). However, BayesMIPCA did not converge on 90% data missingness, so a single imputation method that was similar to BayesMIPCA was used for 90% missingness datasets: single imputation using a regularized iterative principal components analysis (Audigier, Husson, & Josse, 2016). For both imputation methods, the number of principal components was determined with parallel analysis (Horn, 1965).

Statistical Analyses

Analyses consisted of three steps: for each criterion and at each level of data missingness, (a) each model was built using the appropriate training dataset; (b) using test personality data, each model predicted each criterion; and (c) the predictive accuracy of each model was determined by calculating the multiple R value between a model's prediction of a criterion and the actual value of the criterion in the

test data.[3] More details of each technique's procedures are described below.

BISCUIT

BISCUIT was run using the "bestScales" function in the "psych" package (Revelle, 2019, version 1.9.11) of R. BISCUIT was the only technique run on datasets with missing data. To increase the speed of computation, BISCUIT was set to use k-fold cross-validation ($k = 10$) instead of bagging. BISCUIT's tuning parameter, the number of best items, was given the full range of possible values, from 1 item to 135 items. An average model was found for each count of items, using k-fold cross-validation. Across counts of items, and for each criterion and level of missingness in the data, the model with the highest cross-validated multiple R was selected.

Lasso

The lasso was run using the "cv.glmnet" function in the "glmnet" package (Friedman, Hastie, & Tibshirani, 2010) of R. The tuning parameter λ was optimized using the function's default sequence of values. An average model was found for each value of λ using k-fold cross-validation ($k = 10$). For each criterion and level of missingness in the data, the model with the lowest cross-validated error was selected.

Elastic Net

The elastic net was also run using the "cv.glmnet" function. For the tuning parameter λ_2, 11 values were tested, from 0 to 1 in increments of .1. For each value of λ_2, the tuning parameter λ was optimized using the function's default sequence of values. An average model was found for each value of λ_2 using k-fold cross-validation ($k = 10$). Across values of λ_2, and for each criterion and level of data missingness, the model with the lowest cross-validated error was selected.

Random Forest

The random forest was run using the "randomForest" function in the "randomForest" package (Liaw & Wiener, 2002) of R. Forty-five personality items were sampled as candidates for each branch of each tree (which was the default value for the function). There were 100 trees per forest model in order to maintain computational feasibility (i.e., less than 1 week of computation for all random forest models).

Regression

Two regression analyses were used as baselines for typical statistical analyses in personality psychology. One regression technique used the Big Five measures as predictors, while the other used the 27 traits of the SPI-27. These basic regression models did not implement any tuning parameters or resampling procedures. Given the high power of the study, all predicting variables were included in every regression model.

Results

Predictive Accuracy

Predictive accuracy of the techniques in 25 total conditions (5 criteria × 5 levels of data missingness) was calculated with Multiple R and R^2 (R^2 was used to calculate ratios of predictive accuracy between models). The elastic net had the highest predictive accuracy in 13 conditions, BISCUIT in 7 conditions, the lasso in 3 conditions, and regression using the SPI-27 in 2 conditions (Figure 2; Tables 10–12 in ESM 1. For R^2, see Figure 1 in ESM 1). Additionally, the elastic net or lasso had the highest predictive accuracy for all five criteria for the complete, 25%, and 50% data missingness conditions. Models generated by the lasso were, on average, 99.8% as predictive as the elastic net models, which indicated that the predictive accuracies of the elastic net and lasso were functionally equivalent.

For complete data, multiple R effect sizes between the elastic net models and the corresponding criteria were: $R_{Education} = .51$; $R_{Health} = .48$; $R_{BMI} = .43$; $R_{SleepQuality} = .42$; and $R_{SmokingFrequency} = .33$. On average across the five criteria, the random forest was the 3rd most predictive technique for complete data, being 85% as predictive as the elastic net; regression using the SPI-27 (4th) was 81% as predictive; BISCUIT (5th) was 69% as predictive; and regression using the Big Five (last) was 42% as predictive.

One aim of the study was to determine whether BISCUIT, relative to other models, gained an advantage in predictive accuracy as data missingness increased. To assess this question, a ratio was found by dividing the accuracy of each BISCUIT model in each condition by the accuracy of the most predictive model in that condition, and these ratios were averaged for each level of data missingness. Consistent with our hypothesis, each increased level of missingness resulted in an improvement to BISCUIT's average comparative predictive accuracy, up to 75% data missingness: for complete data and 25%, 50%, and 75% data missingness, BISCUIT was, on average, 69%, 74%, 83%, and 100% as predictive as the most predictive model, respectively. In the 75% data missingness condition,

---

[3] Multiple imputation generated 20 datasets for each level of data missingness. For each level of data missingness, twenty models were built using the 20 imputed training datasets, each model was applied to 1 of the 20 imputed test datasets, model fits were determined, and model fits were averaged across the 20 predictions.
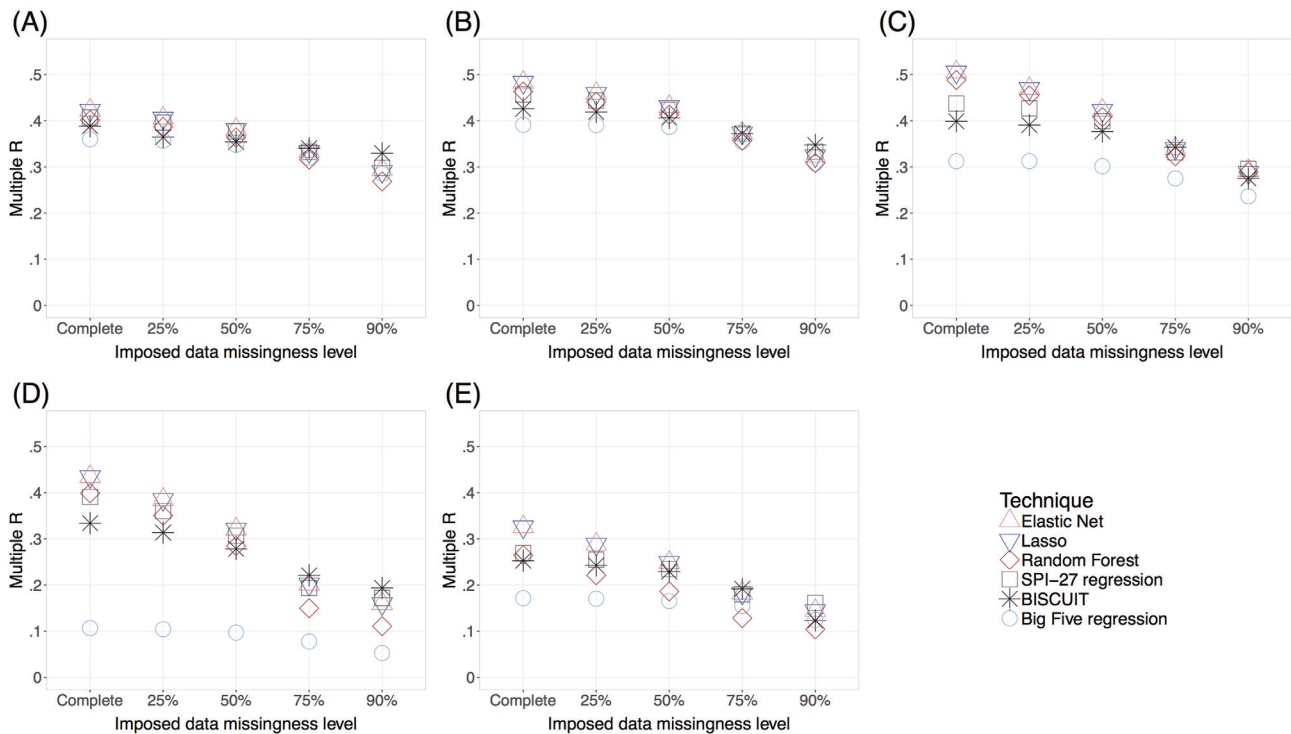
Figure 2. Predictive accuracy (measured in multiple R) of the six statistical techniques, using personality data, across five levels of imposed data missingness, in five criteria. (A) Sleep quality; (B) General heath; (C) Education; (D) Body Mass Index (BMI); (E) Smoking frequency.

BISCUIT had the highest predictive accuracy for four of the five criteria. In the 90% data missingness condition, BISCUIT's comparative predictive accuracy was, on average, 89% as predictive as the most predictive model, and BISCUIT had the highest predictive accuracy for three criteria.[4]

The comparative predictive accuracy of regression using the SPI-27 also improved as data missingness increased: in the 90% data missingness condition, regression using the SPI-27 had the highest predictive accuracy for two criteria. A reviewer was concerned that the superiority of regression using the SPI-27, in the 90% data missingness condition and for the two criteria, was due to regression's tendency to capitalize on chance. They suggested that a model that aggregated regression coefficients across 10 folds would be more stable and less predictive, such that an aggregated regression model using the SPI-27 would not have the highest predictive accuracy for any of the criteria in the 90% data missingness condition. This hypothesis was tested and the results were null: across the five criteria in the 90% data missingness condition, the mean absolute difference in multiple R between the two regression methods was .0008, and the aggregated regression model using the SPI-27 was still the most predictive for the two criteria.

Parsimony

Parsimony of SL models was measured by the number of items used in a model; models that used fewer items were more parsimonious. BISCUIT generated the most parsimonious SL model in 23 of the 25 total conditions (Table 3 in ESM 1).[5] The lasso generated the most parsimonious SL model in 2 of the 25 conditions (Table 4 in ESM 1). SL techniques were ranked for their overall parsimony by calculating the mean and median number of items used in their models across the 25 conditions. Across the 25 conditions, BISCUIT was the most parsimonious technique, using, on average, 30 personality items per model ($Mdn = 30$, $SD = 22$, range = 1–81); the lasso (2nd) used an average of 59

---

[4] We also ran BISCUIT on imputed data to estimate a possible effect of noise generated by imputation. The predictive accuracy of BISCUIT using imputed data was 94% as predictive as BISCUIT using missing data, in terms of R^2 (see Table 13 in ESM 1).

[5] Of note is the fact that BISCUIT generated six 1-item models in the 75% and 90% data missingness conditions. Five of these 1-item models also had the highest predictive accuracy for their condition (Tables 10–12 in ESM 1). See Tables 6–8 in ESM 1 for the item content of three brief BISCUIT models, each predicting a different criterion.

items per model (*Mdn* = 56, *SD* = 27, range = 14–112); the elastic net (3rd) used an average of 60 items per model (*Mdn* = 58, *SD* = 27, range = 15–113; Table 5 in ESM 1); and the random forest (last) used 135 items in every model. The lasso and elastic net used fewer items as missingness increased, whereas the BISCUIT did not.

Post Hoc Analysis

Training Models on Data Missingness Conditions and Testing Them on Complete Data

In the planned analyses, the predictive accuracy of each technique decreased as the amount of data missingness increased (Figure 2). This decrease in predictive accuracy was a combination of two effects: (a) the missingness in the training data, which gave each technique less information with which to build its predictive models; and (b) the missingness in the test data, which gave each technique less information with which to test its predictions. To isolate the first effect, we performed a post hoc analysis to determine the decrease in predictive accuracy of models trained with data missingness but tested on complete data. We selected three techniques: the elastic net, regression using the SPI-27, and BISCUIT. Results indicated that the decrease in predictive accuracy due to missingness in training data was modest (Figure 3; Figure 2 and Tables 15 and 16 in ESM 1). Loss in predictive accuracy was particularly low at the 50% data missingness condition; on average across the five criteria and three techniques, models trained on 50% data missingness were 95% as predictive as their respective models trained on complete data.

SL Techniques on the SPI-27

In the planned analyses, regression using the SPI-27 performed well across missingness levels and criteria. Because SL techniques are supposed to be an improvement over simple regression, we performed a post hoc analysis to determine whether the predictive accuracy of models utilizing the SPI-27 could be improved with either of two SL techniques: the elastic net (the most predictive technique) and BISCUIT (the technique of special interest in this study). Results indicated that the predictive accuracy of models using the SPI-27 was not improved with the use of an SL technique instead of simple regression (Table 14 in ESM 1).

Discussion

BISCUIT

Consistent with our hypothesis, the predictive accuracy of BISCUIT was more competitive with other SL techniques as data missingness increased, up to 75% data missingness,
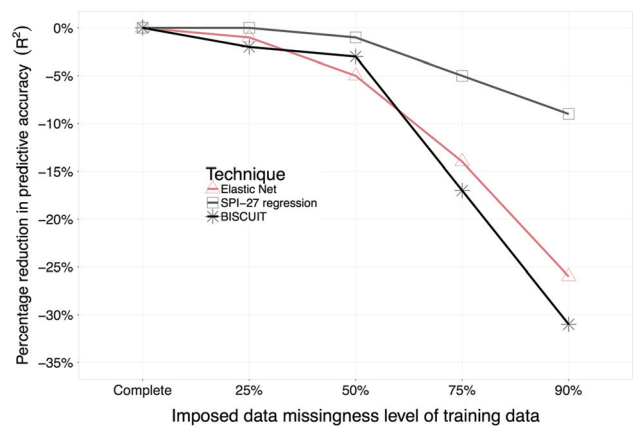

Figure 3. Percentage reduction in predictive accuracy (R^2) for each of three techniques, averaged across five criteria. Each model was trained on one of five levels of imposed data missingness and tested on complete data.

where it generated the model with the highest predictive accuracy in four of five criteria. BISCUIT did not perform as well in the 90% data missingness condition, but it generated the model with the highest predictive accuracy in three of the five criteria. Also consistent with our hypothesis, BISCUIT provided the most parsimonious model in 23 of 25 conditions.

The Elastic Net and Lasso

In terms of predictive accuracy, the elastic net dominated other techniques for the complete data and 25% and 50% data missingness conditions. The lasso was nearly as predictive as the elastic net. The elastic net and lasso may have dominated BISCUIT because BISCUIT's methodology ignored information that the elastic net and lasso did not. Specifically, BISCUIT selected fewer variables than either technique, and BISCUIT used unit-weighting coefficients while the other two techniques used penalized regression coefficients.

The Random Forest

The random forest performed competitively for many missingness conditions and criteria. For complete data, it was 85% as predictive as the elastic net. It is possible that adjusting tuning parameters for the random forest could have increased its predictive accuracy, but we did not find a consensus in the literature regarding what, if any, tuning parameters should be used (Probst & Boulesteix, 2018; Tang et al., 2018). Increasing the number of trees per forest also may have helped, but the random forest was already the most burdensome SL technique in terms of computational load. The random forest appeared to be a lackluster choice for statistical learning with personality

data, due to its suboptimal predictive accuracy, poor parsimony of its models, ambiguities in the literature regarding its tuning parameters, and its burdensome computational load.

Regression Using the SPI-27

Regression using the SPI-27 had greater predictive accuracy than the Big Five (for complete data, it was 93% more predictive), but in most conditions it did not have the maximal predictive accuracy of the elastic net. The SPI-27's dominance over the Big Five is consistent with previous research that found that narrower traits out-predicted broader traits (e.g., Gladstone, Matz, & Lemaire, 2019; Paunonen & Ashton, 2001; Paunonen, Haddock, Forsterling, & Keinonen, 2003). In the 90% data missingness condition, regression using the SPI-27 had the most predictive model for two of five criteria. In such extreme data missingness, the benefit of improving the signal by aggregating items into facet-size factors may outweigh the benefit of utilizing item-level variance in a model's prediction. A post hoc analysis indicated that the predictive accuracy of the SPI-27 was not improved by employing a more complex SL technique instead of simple regression.

Regression Using the Big Five

As expected, regression using the Big Five had poor predictive accuracy compared to other techniques. For complete data, the Big Five was, on average, the least predictive technique of the six tested, being 42% as predictive as the elastic net. In no condition was regression using the Big Five the most predictive model. Additionally, regression using the Big Five showed a relationship between personality and BMI that was far weaker than any other technique (Figure 2; Table 12 of ESM 1). This is consistent with previous findings in which analysis with broader traits failed to find personality-criterion relationships that were evident with narrower traits (Credé, Tynan, & Harms, 2017; Terracciano et al., 2009). If personality researchers continue to use the Big Five to answer the question, "Is personality related to this phenomenon," they may falsely conclude that no relationship exists, when narrower traits would have shown a robust relationship. Thus, regression or correlation using the Big Five may only be appropriate for studying personality-criterion relationships when no alternative is feasible.

Data Missingess

Across all techniques and criteria, predictive accuracy decreased as data missingness increased. However, a post hoc analysis indicated that, after accounting for data missingness in the test data, loss in predictive accuracy was modest. That is, a model trained on a dataset with missing or imputed data is still accurate, but complete data are needed to test this accuracy. Results indicated that the loss in predictive accuracy was approximately 5% for the 50% data missingness condition, which suggests that a large-sample study could introduce 50% data missingness without substantially impacting prediction. Fifty percent data missingness would allow for an item pool twice that of a complete dataset, holding the number of items per participant constant. Ninety percent data missingness would allow for an item pool 10 times that of a complete dataset, but the cost to predictive accuracy would be higher (this study estimated the range of loss to be approximately 10–30%). This loss in predictive accuracy will appear to be even greater if models are not tested on complete data. Thus, whether higher levels of data missingness are optimal for maximizing predictive accuracy will depend on whether the increased predictive accuracy due to a broader item pool will outweigh the loss due to data missingness.

Limitations of the Study

There were at least four methodological decisions that could impact the generalizability of the study's results: First, the comparative predictive accuracy of SL techniques may have depended upon the particular criteria or item pool; new criteria or item pools may favor different SL techniques. Second, only four SL techniques were compared in this study, and only one of them accounted for interactions (the random forest). Other SL techniques, such as Multivariate Adaptive Regression Splines (MARS; Friedman, 1991), may have better accounted for interactions than the random forest did. Third, the criteria chosen in this study were all assumed to be monotonic variables. Results related to the predictive accuracy of BISCUIT cannot be extended to non-monotonic criteria. Fourth, results for this study were based upon MMCAR data and may not generalize to datasets with non-random missingness, such as Missing Not At Random (MNAR) datasets.

Another major limitation of this study is that it compared the predictive accuracy of nuances with higher-order traits using an item pool in which all items were subsumed under higher-order traits. The scales of the SPI-27 (and scales which have followed classic psychometric internal consistency procedures) were designed such that the items were nothing more than representations of a scale; a personality scale does not include items that predict outcomes well but are not exemplars of the scale. Thus, this study may have underestimated the predictive accuracy of nuance-based approaches, given a broader item pool.

Future Directions

Replication and Generalizability of Specific SL Models

Compared to traditional methods of analysis in personality psychology, statistical learning appears to be a more accurate approach to predicting criteria. The success of SL approaches is partially due to modeling the unique variance of personality items, which is ignored in higher-order traits. The superior predictive accuracy of SL techniques seems to suggest that domain-level personality-criterion relationships may be better described as a complex web of nuance-level patterns (e.g., Mõttus, 2016). But how stable are these patterns across datasets? In this study, an elastic net model best predicted BMI in the complete data condition, and this model contained 78 predictors and regression weights. Although the elastic net and other SL techniques did not capitalize on chance fluctuations and outliers, they may have capitalized on idiosyncratic attributes of this dataset. A vital question to answer is: how predictive of a criterion is any specific SL model in a new dataset that has different data collection methods, demographics, or other attributes? Another question to consider is: on average, how similar are two SL models generated from the same technique, using the same pool of predictors, but trained on substantially different datasets? Further research will be required to determine the generalizability of any given SL model, and whether parsimonious SL models are more replicable than complex SL models.

Utilizing a Planned Missing Data Structure to Train Statistical Learning Models

Post hoc analysis indicated that there was relatively low cost to predictive accuracy for models trained on datasets with missingness, compared to models trained on complete data. In the case of 50% data missingness, loss in predictive accuracy was about 5%. This finding suggests that researchers should consider using planned data missingness in their study designs. Randomly sampling items from a pool, instead of administering the same items to every participant, would allow a study to multiply the number of items in its pool while still allowing for the development of robust statistical learning models. In order for a model trained on MMCAR data to have maximal accuracy in predicting a criterion in a new dataset, one would need to collect complete data on the variables that were included in the model. Of the techniques in this study, BISCUIT tended to have the fewest variables in its models, and in some models it had as few as one predictor (Table 3 in ESM 1). Because it is an accurate, parsimonious and cost-effective statistical learning technique, BISCUIT could prove to be especially useful in applying personality-criterion models to real-world predictions of criteria.

Conclusions

Results from this study indicate that statistical learning techniques could prove to be essential in future research of personality-criterion relationships. SL techniques are low-cost tools that increase the predictive power of personality beyond traditional techniques; greater predictive accuracy is achieved by utilizing the same raw data. Since statistical learning methods excel at modeling item-level variance, item pools that contain a broad array of personality nuances may be valued more highly in the future. Planned data missingness designs are suited to meet the need for larger item pools; a study can collect data on an item pool of virtually any size, while still administering a given number of items per participant. Although both SL techniques and planned data missingess are powerful procedures, both can add complexity to a study. Statistical learning techniques such as BISCUIT offer a balanced approach to the study of personality-criterion relationships, by generating parsimonious models that have greater predictive accuracy than traditional methods.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000590

ESM 1. These equations, tables, and figures show useful information, but they were not vital for the manuscript.

References

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79. https://doi.org/10.1214/09-SS054

Audigier, V., Husson, F., & Josse, J. (2014). Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation, 86*, 2140–2156. https://doi.org/10.1080/00949655.2015.1104683

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification, 10*, 5–26. https://doi.org/10.1007/s11634-014-0195-1

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods, 21*, 603–620. https://doi.org/10.1037/met0000088

Condon, D. M. (2018, January 10). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv Preprint.* https://doi.org/10.31234/osf.io/sc4p9

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology, 113*, 492–511. https://doi.org/10.1037/pspp0000102

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*, 1–67. https://doi.org/10.1214/aos/1176347963

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction Data. *Psychological Science, 30*, 1087–1096. https://doi.org/10.1177/0956797619849435

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Hathaway, S. R., & McKinley, J. C. (1942). *Manual for the Minnesota Multiphasic Personality Inventory* [Computer software manual]. Minneapolis, MN: University of Minnesota Press.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*, 95–108. https://doi.org/10.1038/nrg1521

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada* (pp. 278–280). IEEE Computer Society Press. https://doi.org/10.1109/ICDAR.1995.598994

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185. https://doi.org/10.1007/BF02289447

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to statistical learning: With applications in R* (8th ed.). New York, NY: Springer.

Josse, J., & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Societe Francaise de Statistique, 153*, 79–99.

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, 70*, 1–31. https://doi.org/10.18637/jss.v070.i01

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*, 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review, 19*, 97–112. https://doi.org/10.1177/1088868314541857

Mõttus, R. (2016). Towards more rigorous personality trait-outcome research. *European Journal of Personality, 30*, 292–303. https://doi.org/10.1002/per.2041

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D., & Revelle, W. (2017, June 23). Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. https://doi.org/10.31234/osf.io/4q9gv

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112*, 474–490. https://doi.org/10.1037/pspp0000100.supp

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality, 52*, 47–54. https://doi.org/10.1016/j.jrp.2014.07.005

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait heterogeneity in age group differences in personality domains and facets: Implications for the development and coherence of personality traits. *PLoS One, 10*, e0119667. https://doi.org/10.1371/journal.pone.0119667

Mõttus, R., Sinick, J., Terracciano, A., Hřebčková, M., Ando, J., Mortensen, E. L., . . . Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rankorder stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 117*, e35–e50. https://doi.org/10.1037/pspp0000202

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*, 524–539. https://doi.org/10.1037/0022-3514. 81.3.524

Paunonen, S. V., Haddock, G., Forsterling, F., & Keinonen, M. (2003). Broad versus narrow personality measures and the prediction of behaviour across cultures. *European Journal of Personality, 17*, 413–433. https://doi.org/10.1002/per.496

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research, 18*, 1–8. Retrieved from http://jmlr.org/papers/volume18/17-269/17-269.pdf

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research* [Computer software manual]. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *Handbook of online research methods* (pp. 578–596). Thousand Oaks, CA: Sage Publications.

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition* (pp. 27–49). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-1210-7_2

RStudio Team. (2019). *RStudio: Integrated Development Environment for R* [Computer software manual]. Boston, MA: RStudio, Inc.. Retrieved from http://www.rstudio.com/

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics, 6*, Advance online publication. https://doi.org/10.4172/2155-6180.1000224

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality, 32*, 186–201. https://doi.org/10.1002/per.2147

Tang, C., Garreau, D., & von Luxburg, U. (2018, December 3). When do random forests fail? *Conference on Neural Information Processing Systems*. Retrieved from https://dblp.org/rec/conf/nips/TangGL18

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., & Costa, P. T. Jr. (2009). Facets of personality

linked to underweight and overweight. *Psychosomatic Medicine, 71*, 682–689. https://doi.org/10.1097/PSY.0b013e3181a2925b

Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213–217. https://doi.org/10.1037/0033-2909.83.2.213

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika, 73*, 691–703. https://doi.org/10.1007/s11336-008-9066-z

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3*, 23–40. https://doi.org/10.1007/BF02287917

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

ORCID
Lorien G. Elleman
 https://orcid.org/0000-0001-6689-0059

Lorien G. Elleman
Department of Psychology
Northwestern University
Swift Hall 102
2029 Sheridan Road
Evanston, IL 60208
USA
lgelleman@u.northwestern.edu