Received: 15 February 2019 | Revised: 23 November 2019 | Accepted: 28 November 2019

DOI: 10.1111/ijsa.12279

FEATURE ARTICLE

International Journal of SELECTION AND ASSESSMENT | WILEY

Using a supervised machine learning algorithm for detecting faking good in a personality self-report

Pierpaolo Calanna[1] | Marco Lauriola[1] | Aristide Saggino[2] | Marco Tommasi[2] | Sarah Furlan[3]

[1]Department of Dynamic and Clinical Psychology, Sapienza University of Rome, Rome, Italy

[2]Department of Psychological Sciences Humanities and Territory, Gabriele D'Annunzio University of Chieti-Pescara, Chieti, Italy

[3]Giunti Psychometrics, Florence, Italy

Correspondence
Pierpaolo Calanna, Department of Dynamic and Clinical Psychology, Sapienza University of Rome, Via degli Apulii, 1, 00185 Rome, Italy.
Email: pierpaolo.calanna@uniroma1.it

Abstract

We developed a supervised machine learning classifier to identify faking good by analyzing item response patterns of a Big Five personality self-report. We used a between-subject design, dividing participants (N = 548) into two groups and manipulated their faking behavior via instructions given prior to administering the self-report. We implemented a simple classifier based on the Lie scale's cutoff score and several machine learning models fitted either to the personality scale scores or to the items response patterns. Results shown that the best machine learning classifier—based on the XGBoost algorithm and fitted to the item responses—was better at detecting faked profiles than the Lie scale classifier.

KEYWORDS

assessment, measurement, personality, statistics, testing

1 | INTRODUCTION

In the context of personality assessment using self-rating scales, the term *faking good* refers to the deliberate attempts to alter one's responses to represent oneself in a favorable light and attain some personal goal (Ziegler, MacCann, & Roberts, 2011). Field experts use several terms to describe the phenomenon, which represent a broad spectrum of meaning. For example, some researchers consider faking good to be a response style—a disposition that is stable across time and questionnaires (e.g., Helmes & Holden, 1986), whereas others consider it to be a response set—underlying its transient nature and situation-dependence (e.g., Paulhus, 2002). Some scholars view faking good as a response bias—a perturbation of the measurement process that should be controlled for (e.g., van de Mortel, 2008); yet others treat it as a substantive characteristic or personality construct that needs to be accounted for (e.g., McCrae & Costa, 1983).

Not only do definitions vary, but researchers also disagree on the immediate consequences of faking good. Although it has been shown to affect mean scores on nearly all personality dimensions of the Five-Factor Model, with effect sizes ranging from moderate to large (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006), interpretations of the consequences are still contentious. For example, several studies found that faking good deteriorates the factor structure of questionnaires (e.g., Topping & O'Gorman, 1997), but others have not found this (e.g., Ellingson, Smith, & Sackett, 2001). Some authors have provided evidence of a decrease in the criterion validity of personality scales due to faking (e.g., Holden, 2007), but others have failed to replicate such findings (e.g., Kurtz, Tarquini, & Iobst, 2008).

Although there are different perspectives on the nature of faking good and its consequences, researchers and practitioners do agree on the benefits of assessing it in self-report questionnaires (Goffin & Christiansen, 2003). Scales designed to detect dishonest responding are one of the most widespread lying-detection strategies (Paulhus, 1991), and their incorporation into many high-stakes personality inventories is a clear indication of the ubiquitous concern about the impact of faking good. These scales (i.e., Lie scales) usually consist of unrealistic self-serving statements such as '*I always try to practice what I preach*' that people high on deception tend to endorse (Crowne & Marlowe, 1960). The conceptual premise behind Lie scales is that faking good can be approximated to a linear/quasi-linear process and detected by gradually accumulating evidence until some specified

threshold is reached and eventually surpassed. In other words, the individual's responses to the Lie items are summed up, and the total is checked against a normative cutoff; scores above this value indicate significant distortion of the personality profile being assessed.

Two recent studies tried to tackle faking good from a different angle. Kuncel and Borneman (2007), and Kuncel and Tellegen (2009) showed that (a) faking good not only takes place at the level of scale scores, but it also occurs at the level of item scores. Further, (b) its nature can be inherently nonlinear. These two properties translate into *idiosyncratic item responses* that may act as markers of profile distortions, which may be used as an alternative way of detecting faking good, besides the more traditional Lie scales.

Intuitively, analyzing response patterns is a more complex endeavor than analyzing Lie scale scores. It is important to note that a line of research that uses item response theory (IRT) to analyze such patterns has emerged (e.g., Brown & Harvey, 2003; Holden & Book, 2009; Zickar, Gibby, & Robie, 2004). Results thus far are mixed and led some scholars to advise against the use of IRT-based techniques in applied settings (Zickar & Sliter, 2011). Machine learning algorithms seem to be a natural and viable approach to the challenges posed by having to scrutinize item responses. Machine learning algorithms were specifically developed for processing high dimensional, linear/nonlinear data like these (Marsland, 2014). Such approaches have been already adopted in psychology (e.g., Gladstone, Matz, & Lemaire, 2019; Park et al., 2015; Youyou, Kosinski, & Stillwell, 2015), but—to our knowledge—the present study is among the first attempts at using them in the evaluation of faking good, taking a similar approach of other recent papers (e.g., Dua & Bais, 2014; Goerigk et al., 2018).

2 | THE CURRENT STUDY

Inspired by the work Kuncel and Borneman (2007), this exploratory study was aimed at developing a technique for detecting faking good in personality self-reports that could be used as an alternative to the Lie scales. To attain this goal, we posed two questions: Can we employ supervised machine learning algorithms to scrutinize personality scale scores or—at a deeper level—item response patterns to detect faking behaviors? Can machine learning classifiers be used in lieu of the Lie scales?

Machine learning algorithms solve regression and classification problems by generating mathematical models that use a chosen set of variables to predict continuous or categorical target outcomes (Kotsiantis, Zaharakis, & Pintelas, 2006). Generally speaking, a supervised machine learning workflow entails a series of different stages, from the collection and preprocessing of data to the implementation of several competing models that are trained, tuned, and tested through solved cases (i.e., correctly labeled observations acting as meaningful exemplars; Marsland, 2014). In the training stage, the models learn how to map each observation to its corresponding label. This process is not a question of memorizing such associations, but rather one of knowledge extraction via a human-like form of inductive reasoning (Xue & Zhu, 2009). The learning process implies numerical optimization techniques that iteratively update the models to improve their predictive ability. A tuning stage is also needed because machine learning algorithms have configurable settings that influence their generative behavior. Such settings—which are called hyper-parameters and can be thought of as sensitive control knobs—have to be calibrated to squeeze out the best possible performance from the models (Swamynathan, 2017). In essence, the tuning stage corresponds to feeding the algorithms with different sets of hyper-parameters (before the learning process begins) and choosing those that optimize the models being generated and trained. In the final stage of the workflow, the candidate models are tested through some performance indices: the best model is eventually chosen and effectively deployed by applying it to new observations. For classification problems, the most used performance indices are: (a) *Accuracy*—the ratio of correctly classified observations to the number of observations in the data; (b) *Precision*—the ratio of correctly classified positive observations[1] to the number of observations labeled as positive by the model; (c) *Recall*—the ratio of correctly classified positive observations to the number of positive observations in the data; (d) *F*1—the harmonic mean of precision and recall; and (e) *AUC*—the area under the Receiver Operating Characteristic (ROC) curve (Sokolova & Lapalme, 2009). It is important to note that the training, tuning, and testing stages should be performed with independent data. To do otherwise is to risk an overly optimistic estimate of model performance (Luo et al., 2016). One possible technique to tackle the issue is *k-fold cross-validation* (Hastie, Tibshirani, & Friedman, 2001), wherein the available data (i.e., the collection of solved cases) are randomly divided into *k* subsets or folds; the model being developed is trained/tuned on *k*-1 folds and then tested on the left-out one. This process is repeated *k* times, and the model's performance across the *k* rounds is averaged, resulting in more stable estimates.

In our exploratory study, we implemented several machine learning classifiers on an existing online repository of honest versus faked profiles derived from a widely used personality self-report. These classifiers were fitted either to the scale scores or the item response patterns to choose the most performant classifier. We then implemented a between-subject design by administering the questionnaire to the participant sample with different standard/fake-inducing instructions. Finally, we deployed two predictive models: a simple classifier based upon the questionnaire Lie scale—serving as a benchmark—and the best-implemented machine learning classifier. We intended to determine whether the latter was at least as effective as the benchmark model at detecting distorted profiles.

Concerning the choice of machine learning algorithms, we opted for the ensemble methods, which are a subset of supervised learning techniques. Instead of relying on a "monolithic" predictive model, ensemble methods combine several base-learner units to provide a robust committee of estimators. The key point is that each unit of the committee has a diverse, decentralized, and relatively independent knowledge of the problem and contributes via an aggregation mechanism to a better composite prediction (Rokach, 2010).

From the available range of ensemble methods, we selected random forests and XGBoost. The choice was made following Hastie et al.'s (2001) observation that decision trees—and extensively tree-based ensemble algorithms—are a good option for an *off-the-shelf* data mining strategy, which is defined as: "one that can be directly applied to the data without requiring a great deal of time-consuming data preprocessing or careful tuning of the learning procedure". (p. 350). They are quick to implement, robust to outliers and missing data in the inputs, can naturally deal with continuous and categorical variables, perform implicit variable selection, and can capture non-linear relationships (Hastie et al., 2001).

Tree-based ensemble algorithms employ different strategies to create the collection of decision trees. Random forests, for example, use bootstrapped samples in the tree-building process and inject a further quantum of randomness by arbitrarily choosing a subset of predictors to split each node of every tree. Such forced randomness improves the ensemble accuracy while safeguarding its generalizability (Breiman, 2017). XGBoost, on the other hand, adds trees in a stepwise manner: each new tree tries to correct its predecessors' mistakes (Géron, 2019). The effectiveness of this iterative process relies on XGBoost learning objective function, which encompasses a loss term and a regularization term. The former measures the difference between predictions and the ground truth (i.e., accuracy), while the latter penalizes overly complex solutions (i.e., parsimony). When correctly optimized, XGBoost learning objective function ensures high prediction hits and promotes generalizability (Chen & Guestrin, 2016). Apart from random forests and XGBoost, we also employed a logistic regression because it is commonly used in binary classification tasks (Perlich, Provost, & Simonoff, 2003). In the machine learning paradigm, logistic regression models are iteratively optimized through a regularized loss function to find the hyperplane that best separates one class—assumed to be the target—from the other; this hyperplane is a linear combination of the given predictors (Dreiseitl & Ohno-Machado, 2002).

3 | METHODS

3.1 | Participants

Participants were 548 undergraduate psychology students from "Sapienza" University of Rome and "G. D'Annunzio" University of Chieti-Pescara (M_{age} = 22.11, *SD* = 3.45). They completed a self-report personality questionnaire based on three sets of instructions (honest respondents, fake teachers, fake firefighters).

In addition to the Participant Sample, data were collected from a repository of real-world online assessments, which occurred prior to the present study, and were comprised of 4,000 cases (M_{age} = 32.27, *SD* = 8.95), 2,000 of which were labeled as *High Fakers* (Lie scale: *M* = 3.30, *SD* = 0.39) and the rest as *Low Fakers* (Lie scale: *M* = 2.29, *SD* = 0.33).

Tables 1 and 2 summarize the gender composition of both samples.

TABLE 1 Gender composition of participant dataset

	Condition		
Gender	Fake	Honest	All
Female	187	220	407
Male	67	74	141
All	254	294	548

Note: Fake = Fake teachers (54%) + Fake firefighters (46%).

TABLE 2 Gender composition of online repository

	Condition		
Gender	Fake	Honest	All
Female	830	816	1646
Male	1,170	1,184	2,354
All	2,000	2,000	4,000

The local ethical review board at "Sapienza" University of Rome approved the study.

3.2 | Measures

We set two major requisites for the personality self-report to be employed in this exploratory study: (a) it had to be an assessment tool commonly adopted by Italian Human Resources professionals; and (b) it had to have a Lie scale for the detection of faking. In light of these considerations, we chose the Big Five Questionnaire-2 (BFQ2; Caprara, Barbaranelli, Borgogni, & Vecchione, 2007), a 134-item self-report with domain scales closely reflecting the Five-Factor Model: Extraversion, Agreeableness, Conscientiousness, Emotional Stability (inverse of Neuroticism), and Openness. Respondents use a five-point Likert scale ranging from *not at all true* to *completely true* to indicate the extent to which each item describes their personal experience. The BFQ2 has a Lie scale to detect self-enhancing/distorting strategies. It consists of 14 items which—when endorsed—depict the unrealistic picture of an individual who is capable, competent, brilliant, courageous, particularly affable, respectful of others, and attentive to social norms (Caprara et al., 2007). The Lie scale has a normative gender-based cut-off score of 55T, a value above which faking behaviors are considered meaningful, spanning from moderate to marked levels.

3.3 | Procedure

The first part of the procedure involved training, tuning, and testing several machine learning classifiers using the online repository, either fitting them to the personality scale scores or the item response patterns (Table 3). Our intention was to disentangle the effect of using different predictors from the effect of implementing different algorithms.

TABLE 3 Implemented machine learning classifiers

Model name	Algorithm	Predictors
LR-S	Logistic regression	Scale scores (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness)
RF-S	Random forest	
XGB-S	XGBoost	
LR-I	Logistic regression	Item response patterns (All items except Lie scale items)
RF-I	Random forest	
XGB-I	XGBoost	

In the second part of the procedure, we compared the performance of the best machine learning classifier from the previous step with a benchmark model based on the BFQ2 Lie scale (set up using gender-specific cutoff scores). Both classifiers were then applied to 1,000 bootstrapped replicates of the participant sample which was collected by administering the BFQ2 under two different conditions: (a) the honest group completed the questionnaire after receiving standard instructions; (b) participants in the fake group were asked to imagine they were candidates for a job either as a high school teacher or as a firefighter and to complete the questionnaire in a way that would enhance their chance of being selected (Appendix A in Supporting Information).

3.4 | Analysis

Statistical computations where performed: (a) with STATA (ver. 14.2); (b) with the Python language (ver. 3.7.3) using the following libraries: Scipy (ver. 1.3.0), Numpy (ver. 1.16.4), Pandas (ver. 0.24.2), Matplotlib (ver. 3.1.0), Scikit-learn (ver. 0.21.2), XGBoost (ver. 0.90); (c) with the R language (ver. 3.5.2) using the following packages: XGboost (ver. 0.82.1), ALEPlot (ver. 1.1).

4 | RESULTS

As already stated, we trained, tuned, and tested several machine learning classifiers on the online repository, either fitting them to the BFQ2 personality scale scores or the item response patterns. To safeguard results' generalizability, we adopted a multi-layer cross-validation procedure. More precisely, we ran (a) a 5-fold cross-validation loop for tuning our algorithms and (b) another 10-fold cross-validation loop for estimating their performance (Cawley & Talbot, 2010). For the hyper-parameters tuning, one may tackle such optimization task with a deterministic approach by systematically scanning the entire hyper-parameters space for the best possible configuration or with a stochastic approach by iteratively selecting random combinations of hyper-parameters with the objective of choosing the most suitable ones (Bergstra, Bardenet, Bengio, & Kégl, 2011). We followed the second approach, and for each model, we used a random search to find the best hyper-parameters relative to the *F*1 score (thus maximizing the tradeoff between precision and recall).

Table 4 reports the accuracy indices of all classifiers. On a predictor-centric level, it is worth noting that models based on item patterns showed consistently better performance than those fitted to scale scores. On an algorithm-centric level, the best classifier was XGB-I. It should be observed that LR-I—the second best classifier—came very close, but in the machine learning realm, it is customary to choose the final model however small the improvement in performance might be (Alpaydin, 2010).

To gain an initial understanding of the inner mechanics of XGB-I, we considered all of its predictors in order of "gain" (i.e., the improvement in accuracy brought by each predictor to the model) and considered the 10 most influential ones: 50% of such predictors belonged to Emotional Stability while the rest were unevenly distributed between Agreeableness (20%), Extraversion (10%), Conscientiousness (10%), and Openness (10%). All of them had a quasi-linear effect on XGB-I predictions as it can be readily seen in

TABLE 4 Performance evaluation of the machine learning classifiers (online repository)

Model name	Machine learning algorithm	Predictors	Mean value over 10-fold Cross-Validation				
			Acc.	Prec.	Rec.	*F*1	AUC
LRC-S	Logistic regression	Scale scores	0.66	0.66	0.67	0.67	0.72
RF-S	Random forest		0.65	0.65	0.63	0.64	0.70
XGB-S	XGBoost		0.65	0.65	0.64	0.65	0.70
LRC-I	Logistic regression	Response patterns	0.76	0.75	0.76	0.76	0.83
RF-I	Random forest		0.74	0.76	0.72	0.74	0.82
XGB-I	XGBoost		0.76	0.76	0.77	0.77	0.84

Note: A dummy model predicting the most frequent class had a mean accuracy of 50%.

Abbreviations: Acc., Accuracy; Prec., Precision; Rec., Recall. *F*1, *F*1 score; AUC, Area under the ROC curve.

their Accumulated Local Effects plots (Appendix B in Supporting Information). These plots describe how single predictors influence model predictions (Apley & Zhu, 2016).

The objective of the last stage of our analysis was to compare XGB-I with the Lie scale classifier taken as a benchmark. To do so, we administered the BFQ2 to the participants of our study giving them different faking instruction sets (see Procedure). Table 5 shows the descriptive statistics of the BFQ2 scales across the three conditions, and Figure 1 depicts the mean profiles associated with each condition. Both faking groups scored higher than honest respondents, although fake firefighters presented less distorted profiles than fake teachers. We conducted a MANOVA to ascertain whether the upward shifts were statistically significant. Pillai's trace = 0.55 indicated that there was an effect of group membership on the BFQ2 scale scores, $F(2, 545) = 34.44$, $p < .001$.

Follow-up ANOVAs with effect sizes and contrasts are shown in Table 6. The omnibus *F*-tests were all significant with effect sizes being: (a) large for Emotional Stability, Lie, Openness and Extraversion, and (b) close to the large threshold for the remaining scales (Kirk, 1996). Contrasts confirmed the pronounced differences between honest respondents and the two faking conditions combined. Furthermore, there were significant differences between fake teachers and fake firefighters on all scales but Lie.

We also measured to what extent the BFQ2 scale score distributions overlapped between honest and faking groups. Table 7 reports the Szymkiewicz–Simpson coefficient for each pair of distributions, an easy to compute similarity index that expresses the distributional overlap with values ranging from 0 to 1 (i.e., from nonexistent to perfect overlap; Vijaymeena & Kavitha, 2016). Figure 2 provides a visual sense of the matter. Scales showed various degrees of overlap with Emotional Stability having the lowest degree. It is readily apparent that any attempt at separating the groups with a cutoff score would have produced a large number of false positives/negatives across nearly all scales.

We then applied the Lie scale classifier and XGB-I to assign group membership in 1,000 bootstrapped replicates of the participant sample. Results (see Table 8 and Appendix C in Supporting Information) indicated that both the Lie scale classifier and XGB-I performed better than a dummy model predicting the most frequent class. Such a dummy

TABLE 5 Descriptive statistics of BFQ2 scales across conditions (participant sample)

	Honest respondents				Fake teachers				Fake firefighters			
Scale	*M*	*SD*	*SK*	*K*	*M*	*SD*	*SK*	*K*	*M*	*SD*	*SK*	*K*
Extraversion	3.18	0.55	−0.08	−0.25	3.67	0.40	−0.09	0.32	3.37	0.47	−0.24	0.14
Agreeableness	4.04	0.38	−0.66	1.20	4.34	0.36	−1.99	9.54	4.22	0.40	−1.05	2.66
Conscientiousness	3.75	0.48	−0.04	−0.11	4.14	0.36	−0.21	−0.49	4.01	0.40	−0.17	−0.49
Emotional stability	2.75	0.71	0.14	−0.42	3.95	0.41	−0.10	−0.43	3.68	0.50	0.46	−0.76
Openness	3.78	0.48	−0.30	−0.17	4.26	0.38	−0.66	0.39	3.94	0.42	−0.13	−0.01
Lie	2.63	0.53	0.07	−0.10	3.34	0.53	−0.29	−0.12	3.29	0.54	−0.13	−0.05

Note: Abbreviations: *M*, Mean; *SD*, Standard Deviation; *SK*, Skewness; *K*, Kurtosis.

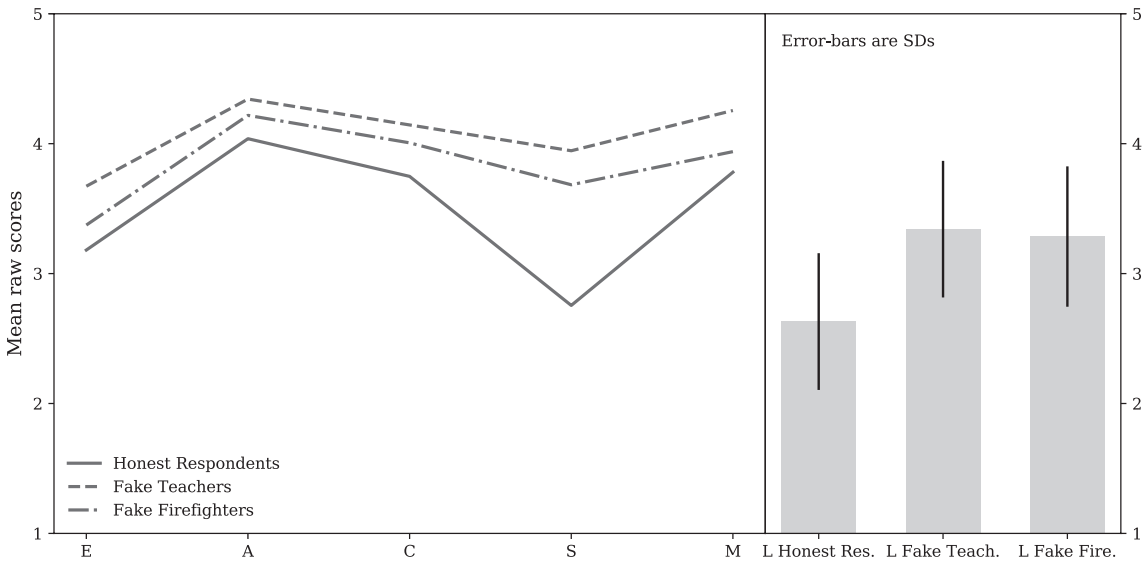

FIGURE 1 BFQ2 scales profiles. E = Extraversion, A = Agreeableness, C = Conscientiousness, S = Emotional Stability, M = Openness, L = Lie

TABLE 6 Descriptive ANOVAs, effect sizes (ω^2) and contrasts for scale scores (participant sample)

BFQ2 Scale	F (2, 545)	ω^2	Contrasts	
			Honest respondents versus Fakers	Fake teachers versus Fake firefighters
Extraversion	45.18 **	0.14	−0.34**	0.30**
Agreeableness	32.25 **	0.10	−0.24**	0.13*
Conscientiousness	43.27 **	0.13	−0.33**	0.14*
Emotional Stability	219.8 **	0.44	−1.06**	0.26**
Openness	53.54 **	0.16	−0.31**	0.32**
Lie	114.6 **	0.29	−0.68**	0.06

Note: Omega squared interpretation: $\omega^2 \geq 0.01$, $\omega^2 \geq 0.06$, and $\omega^2 \geq 0.14$ represent small, medium, and large effect sizes (Kirk, 1996).

*$p < .01$; **$p < .001$.

TABLE 7 Descriptive Szymkiewicz–Simpson coefficient of BFQ2 scales (participant sample)

BFQ2 scale	Honest respondents versus Fakers	Honest respondents versus Fake teachers	Honest respondents versus Fake firefighters
Extraversion	0.70	0.53	0.77
Agreeableness	0.72	0.66	0.79
Conscientiousness	0.70	0.70	0.77
Emotional stability	0.38	0.25	0.45
Openness	0.70	0.51	0.85
Lie	0.54	0.59	0.56

Note: Szymkiewicz–Simpson Coefficient spans from 0 (no overlap) to 1 (complete overlap).

model had a mean accuracy of 54% over the bootstrapped replicates, a value below that of the Lie scale classifier (72%) and XGB-I (82%).

A series of ANOVAs (Table 9) confirmed that XGB-I performed significantly better than the Lie scale model with large effect sizes on all indices (Kirk, 1996).

To gain insight into misclassifications, we also confronted the mean profiles of misclassified cases (grouped by classifier) to the mean profile of the class to which they should have been assigned. This comparison was made by visual inspection (Figure 3) and by calculating the Mahalanobis distance as a proximity measure (Table 10).

Results showed that individuals misclassified by XGB-I had a mean personality profile more dissimilar from the class they belonged to when compared with the mean profile of the benchmark model misclassified cases.

5 | DISCUSSION

At the start of this exploratory study, we asked whether supervised machine learning algorithms could be used to scrutinize item response patterns to detect faking behaviors in a personality questionnaire. The idea was inspired by the seminal work of Kuncel and colleagues (2007, 2009), who suggested that a fine-grained analysis of item responses could be effective in detecting distorted profiles of personality self-reports.

On this basis, we implemented three classifiers—a logistic regression, a random forest, and an XGBoost machine—all of which were trained with an online repository of real-world assessments, divided into honest and faked profiles. The machine learning algorithms were fitted either to the personality scale scores (LR-S, RF-S, XGB-S) or the item response patterns (LR-I, RF-I, XGB-I) for comparison purposes. We found that using response patterns instead of scale scores increased the accuracy of all classifiers. This result offered an affirmative answer to the first research question and confirmed the suggestion that the faking problem could be addressed by scrutinizing item patterns (Kuncel & Borneman, 2007). Our results contribute to existing knowledge by demonstrating that faking can be detected at the item level with machine learning classifiers. Intuitively, one would expect the subtle relationships between responses to hold information about faking that is not exploited by detection strategies based solely on scale scores.

Among the tested classifiers, XGB-I turned out to be the best one. Inspection of the list of its most influential predictors revealed that the largest group consisted of items related to Emotional Stability followed by Agreeableness, Extraversion, Conscientiousness, and Openness. All the predictors had a quasi-linear influence over XGB-I performance, as the ALE plots showed. The second most effective classifier was LR-I, a logistic regression fitted to the personality items. The differences between LR-I and XGB-I were negligible. These results suggest that the problem of separating honest respondents from fakers can often be tackled using a linear approach (i.e., LR-I) with minimal loss in accuracy of predictions, although tree-based algorithms would be better at detecting whatever nonlinearity with which the response patterns might be imbued (Hastie et al., 2001).

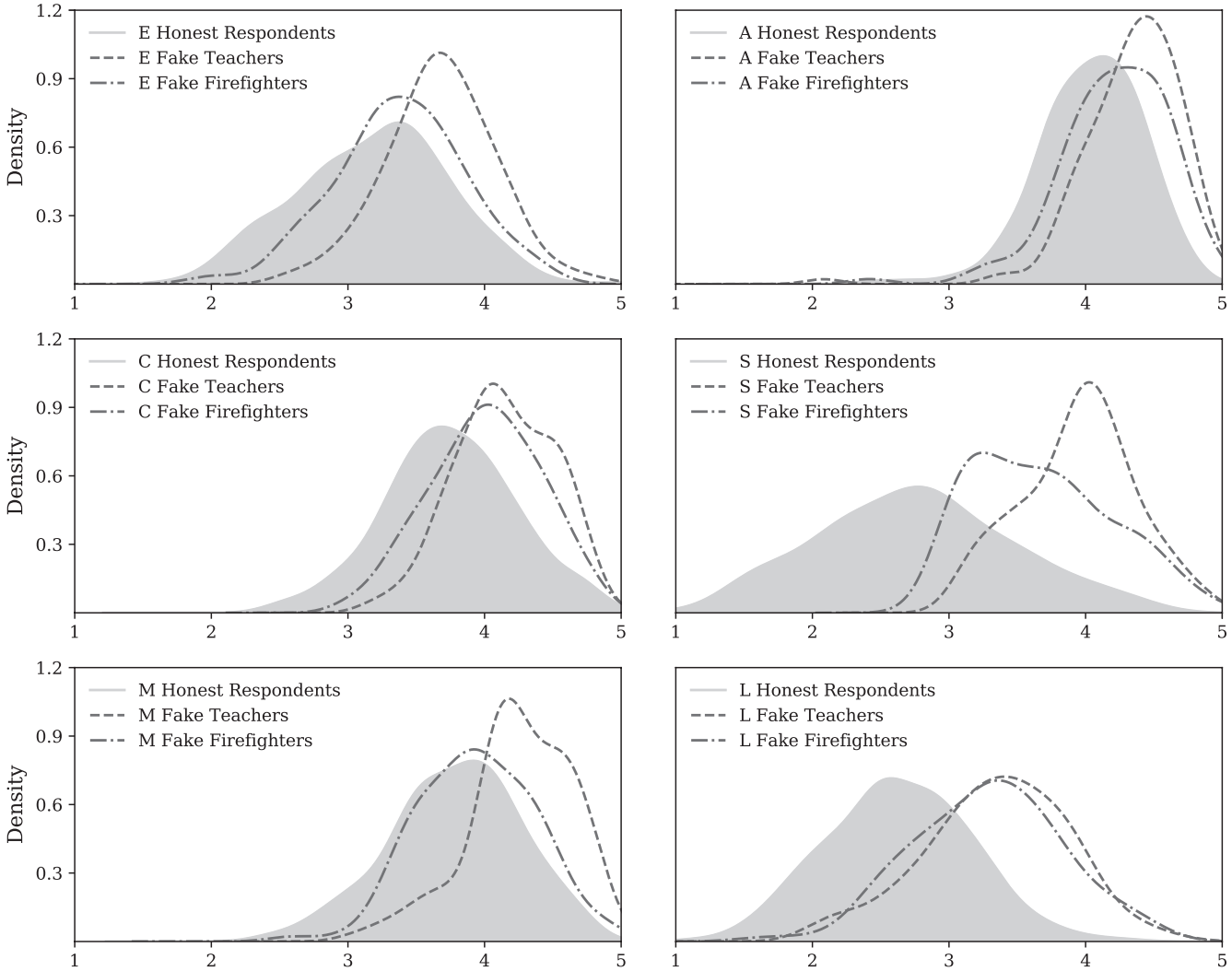

FIGURE 2 BFQ2 scale scores distributions. E = Extraversion, A = Agreeableness, C = Conscientiousness, S = Emotional Stability, M = Openness, L = Lie

Bootstrap replicates	Model	Acc.	Prec.	Rec.	*F1*	AUC
Honest respondents versus Fakers	Lie scale model	0.72	0.72	0.65	0.68	0.72
	XGB-I	0.82	0.82	0.79	0.80	0.90
Honest respondents versus Fake teachers	Lie scale model	0.68	0.75	0.66	0.70	0.69
	XGB-I	0.83	0.86	0.83	0.84	0.91
Honest respondents versus Fake firemen	Lie scale model	0.75	0.68	0.63	0.65	0.73
	XGB-I	0.82	0.77	0.76	0.77	0.88

TABLE 8 Descriptive performance evaluation of lie scale model versus XGB-I (1,000 bootstrap replicates of participant sample)

Note: A dummy model predicting the most frequent class had a mean accuracy of 54%.

Abbreviations: Acc., Accuracy; Prec., Precision; Rec., Recall; *F1*, *F1* score; AUC, area under the ROC curve; XGB-I, XGBoost fitted to item response patterns.

Because professional psychologists typically assess faking behaviors with the aid of Lie scale scores in high-stake situations (Paulhus, 1991), we compared the performance of our best machine learning classifier with a benchmark model based upon the BFQ2 Lie scale. To make this final comparison, we collected new data from undergraduate psychology students (participant sample) whose responses were manipulated by giving them different instructions before administering the BFQ2. In line with the literature (e.g., Birkeland et al., 2006), the faking groups had higher scores—with medium to large effect sizes—than the honest group for all the BFQ2 scales. Noticeably, the Emotional Stability effect size was larger than that for the Lie scale, that is fakers arranged

their self-promoting strategy by endorsing Emotional Stability items more than other items, even those items devised to detect response distortions. Nonetheless, honest and faking groups substantially overlapped on nearly all the scale distributions, suggesting that it could be challenging to separate them using some scale-level cutoff score.

XGB-I and the Lie scale model were then applied to 1,000 bootstrapped replicates of the participant sample. Results showed that XGB-I was consistently more accurate than the Lie scale classifier, allowing us to provide an affirmative answer to the second research question: machine learning classifiers can indeed be used in place of Lie scale scores to reveal distorted profiles. When the performance was evaluated in terms of misclassifications, XGB-I continued to outcompete the benchmark model. Honest respondents and fakers misclassified by XGB-I had personality profiles more dissimilar to the prototypical profile of the group to which they belonged than the respondents misclassified by the Lie scale model. This finding suggests that the type of errors made by XGB-I was subtler, and perhaps more excusable.

TABLE 9 Descriptive ANOVAs, effect sizes of performance indices between the lie scale classifier and XGB-I over 1,000 bootstrap replicates of participant sample (Honest respondents versus Fakers)

Scoring metrics	*F* (1, 1999)	ω^2
Accuracy	17,592.26*	0.90
Precision	6,778.66*	0.77
Recall	14,718.33*	0.88
*F*1	16,687.82*	0.89
AUC	62,021.46*	0.97

Note: XGB-I = XGBoost fitted to item response patterns. Omega squared interpretation: $\omega^2 \geq 0.01$, $\omega^2 \geq 0.06$, and $\omega^2 \geq 0.14$ represent small, medium, and large effect sizes (Kirk, 1996).
*$p < .001$.

The present exploratory study has two limitations. First, the direct manipulation of faking behaviors has been criticized by some authors. For example, Griffith and Peterson (2006) noted that direct faking studies could, at most, provide an estimate of the upper limit of faking but cannot give a clear indication of the extent to which individuals are motivated to fake in real-world settings. An implicit confirmation of the problem comes from the fact that the machine learning classifiers made better predictions with the participant sample—whose faking behaviors were "strained" by direct manipulation—rather than with the online repository of real-world assessments. In any case, all machine learning classifiers were consistently more performant than the Lie scale benchmark model. Future studies should address this issue by assessing individuals who are motivated to fake in real scenarios and investigating whether machine learning classifiers would still be able to differentiate them from honest individuals taken as controls.

Second, the data set used for the comparison of XGB-I and the Lie scale classifier (participant sample) was gender-skewed, with young women being over-represented, and had a narrow age range. We did not appraise XGB-I accuracy with a more heterogeneous data set. It is worth noting that, from another perspective, this limitation could be considered as evidence that the machine learning classifier has good generalizability, as XGB-I was trained on the online repository, which had a different demographic signature.

Apart from the above mentioned limitations, the approach proposed here would allow the length of personality inventories to be reduced by eliminating the control scale items, as XGB-I did not rely on them to make its predictions. It is worth noting that

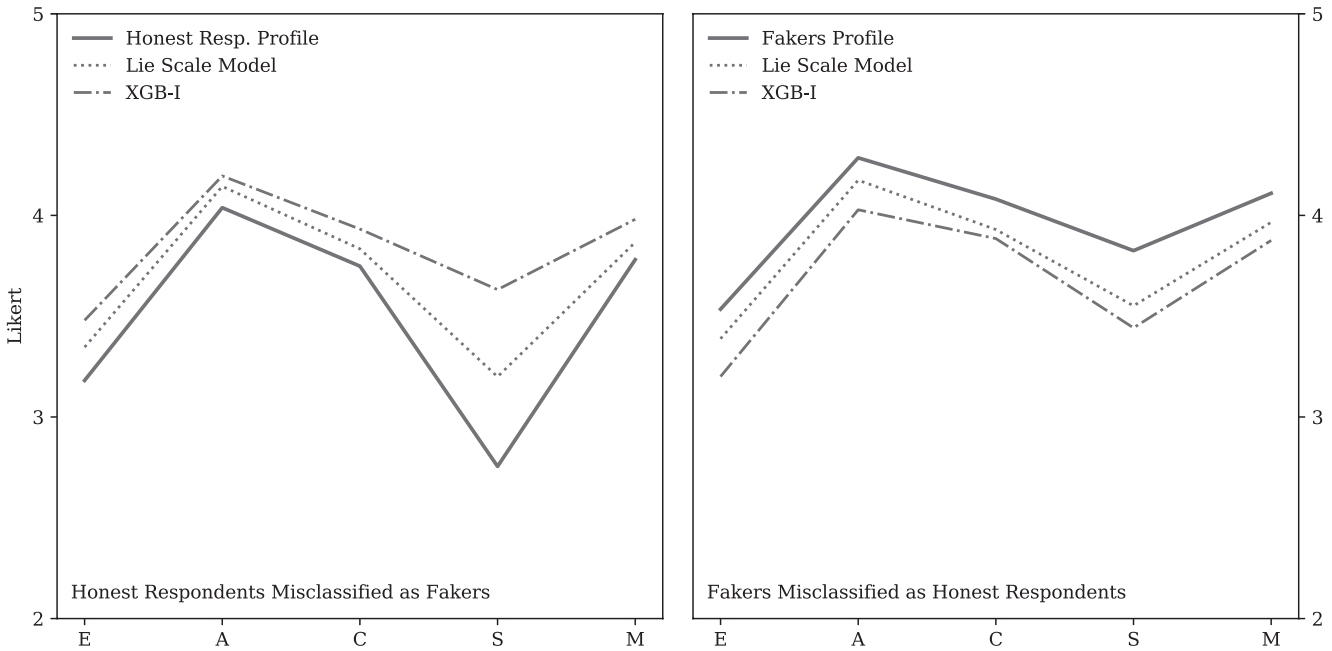

FIGURE 3 Profiles of misclassified cases versus True mean profiles. E = Extraversion, A = Agreeableness, C = Conscientiousness, S = Emotional Stability, M = Openness

TABLE 10 Descriptive Mahalanobis distance of misclassified cases from true group (participant sample)

	Mahalnobis distance from true group	
Misclassifications	**Lie scale model**	**XGB-I**
Honest respondents misclassified as Fakers	0.72	1.38
Fakers misclassified as Honest Respondents	0.64	1.02

Note: XGB-I = XGBoost fitted to item response patterns.

our approach could be used for adding a faking good detection mechanism to those self-reports that do not include a Lie scale. In summary, we feel we have collected enough evidence concerning the BFQ2 to claim that machine learning classifiers exploiting the subtle information hidden in item response patterns represent a promising alternative to Lie scales when it comes to detecting faking good.

ORCID

Pierpaolo Calanna https://orcid.org/0000-0002-4901-3393
Marco Lauriola https://orcid.org/0000-0003-3996-9567
Aristide Saggino https://orcid.org/0000-0002-4903-9833
Marco Tommasi https://orcid.org/0000-0002-4876-0530
Sarah Furlan https://orcid.org/0000-0002-7706-6150

ENDNOTE

[1]Positive observations are those belonging to the target class we wish to predict.

REFERENCES

Alpaydin, E. (2010). *Introduction to machine learning: Third edition*. Cambridge, MA: MIT Press.

Apley, D. W., & Zhu, J. (2016). *Visualizing the effects of predictor variables in Black Box Supervised Learning Models*. Retrieved from http://arxiv.org/abs/1612.08468

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 2546–2554). Red Hook, NY: Curran Associates Inc.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Breiman, L. (2017). *Classification and regression trees*. Wadsworth, CA: Routledge.

Brown, R. D., & Harvey, R. J. (2003). Detecting personality test faking with appropriateness measurement: Fact or fantasy. Annual Conference of the Society for Industrial and Organizational Psychology. Retrieved from http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Vecchione, M. (2007). *Big five questionnaire: Manual*. Firenze, Italy: Organizzazioni Speciali.

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

Chen, T., & Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco, CA: ACM.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Dua, P., & Bais, S. (2014). Supervised learning methods for fraud detection in healthcare insurance. *Intelligent Systems Reference Library*, *56*, 261–285. https://doi.org/10.1007/978-3-642-40017-9_12

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86*(1), 122–133. https://doi.org/10.1037/0021-9010.86.1.122

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media Inc.

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science*, *30*(7), 1087–1096. https://doi.org/10.1177/0956797619849435

Goerigk, S., Hilbert, S., Jobst, A., Falkai, P., Bühner, M., Stachl, C., … Sarubin, N. (2018). Predicting instructed simulation and dissimulation when screening for depressive symptoms. *European Archives of Psychiatry and Clinical Neuroscience*, 1–16. https://doi.org/10.1007/s00406-018-0967-2

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment*, *11*(4), 340–344. https://doi.org/10.1111/j.0965-075X.2003.00256.x

Griffith, R. L., & Peterson, M. H. (2006). *A closer examination of applicant faking behavior*. Charlotte, NC: Information Age Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer.

Helmes, E., & Holden, R. R. (1986). Response styles and faking on the basic personality inventory. *Journal of Consulting and Clinical Psychology*, *54*(6), 853–859. https://doi.org/10.1037/0022-006X.54.6.853

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science*, *39*(3), 184–201. https://doi.org/10.1037/cjbs2007015

Holden, R. R., & Book, A. S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences*, *47*(3), 185–190. https://doi.org/10.1016/j.paid.2009.02.024

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 156–190. https://doi.org/10.1007/s10462-007-9052-3

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*(2), 220–231. https://doi.org/10.1111/j.1468-2389.2007.00383.x

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of ttems: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences*, *45*(1), 22–27. https://doi.org/10.1016/j.paid.2008.02.012

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., … Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, *18*(12), e323. https://doi.org/10.2196/jmir.5870

Marsland, S. (2014). *Machine learning: An algorithmic perspective*. Boca Raton, FL: Chapman and Hall/CRC.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*(6), 882–888. https://doi.org/10.1037/0022-006X.51.6.882

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In B. I. Henry, J. N. Douglas, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (Vol. *4*, pp. 49–69). Mahwah, NJ: Erlbaum Associates.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, *4*(6), 211–255. https://doi.org/10.1162/153244304322972694

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Swamynathan, M. (2017). *Mastering machine learning with Python in six steps*. New York, NY: Apress.

Topping, G. D., & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences*, *23*(1), 117–124. https://doi.org/10.1016/S0191-8869(97)00006-8

van de Mortel, T. F. (2008). Faking it: Social desirability response bias in selfreport research. *Australian Journal of Advanced Nursing*, *25*(4), 40–48.

Vijaymeena, M., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*(1), 19–28. https://doi.org/10.5121/mlaij.2016.3103

Xue, M., & Zhu, C. (2009). *A study and application on machine learning of artificial intellligence*. IJCAI International Joint Conference on Artificial Intelligence, Hainan Island, China, 272–274. https://doi.org/10.1109/JCAI.2009.55

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 of the United States of America*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An Application of Mixed-Model Item Response Theory. *Organizational Research Methods*, *7*(2), 168–190. https://doi.org/10.1177/1094428104263674

Zickar, M. J., & Sliter, K. A. (2011). Searching for unicorns: Item response theory-based solutions to the faking problem. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 113–130). Oxford, UK: Oxford University Press.

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). *New perspectives on faking in personality assessment*. Oxford, UK: Oxford University Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Calanna P, Lauriola M, Saggino A, Tommasi M, Furlan S. Using a supervised machine learning algorithm for detecting faking good in a personality self-report. *Int J Select Assess*. 2020;28:176–185. https://doi.org/10.1111/ijsa.12279