Machine Learning can Detect Faking on Self-Reports and Implicit Association Tests (IATs)

Jessica Röhner, Philipp Thoss, & Astrid Schütz
University of Bamberg, Germany

10.20378/irb-104669

aps 2021 VIRTUAL CONVENTION MAY 26-27, 2021

RESEARCH OBJECTIVE

Theoretical Background

- Even experts cannot detect faking above chance (Fiedler & Bluemke, 2005)
- Recent studies (Boldt et al., 2018; Calanna et al., 2020) have suggested that machine learning may help in this endeavor
- The ability of classifiers to detect faking depends on which classifiers are implemented (logistic regression vs. random forest vs. XGBoost; Calanna et al., 2020)
- The ability of classifiers to detect faking also depends on the type of input data (response patterns vs. scores; Calanna et al., 2020)

⇒However, faking differs with respect to faking conditions, and previous efforts have not taken these differences into account

Shortcomings and Open Questions

1. Faking differs between measures (e.g., faking is more pronounced for self-reports than for objective or indirect measures, such as IATs, Röhner et al., 2011; see also Ziegler et al., 2007)
2. Faking might depend on the construct that is being faked (e.g., Steffens, 2004)
3. Faking depends on the faking direction (faking good vs. faking bad, Bensch et al., 2019; faking high scores vs. low scores, Röhner et al., 2013)
4. Faking depends on whether people have knowledge about the measurement procedure and whether they are provided with strategies on how to fake (i.e., informed faking) or not (i.e., naïve faking; Röhner et al., 2013)
5. Faking is more pronounced when participants are able to practice faking beforehand (Röhner et al., 2011)

- Also, previous efforts did not include empirically supported faking indices in machine learning input data

Hypotheses

1. Given that faking is stronger on self-reports than on IATs, can classifiers spot fakers better on self-reports than on IATs?
2. Considering that constructs might be fakeable to different extents, can classifiers spot fakers comparably well on conscientiousness, extraversion, need for cognition, and self-esteem?
3. Given that faking low is more pronounced than faking high, are classifiers better at spotting the faking of high scores?
4. Given that informed faking is stronger than naïve faking, are classifiers better at spotting informed faking?
5. Given that faking gets somewhat stronger with practice, are classifiers better at detecting faking from experienced fakers than unexperienced fakers?
6. Concerning self-reports, can we replicate the finding that response patterns are better than scores as input data for machine learning in faking detection? For IATs, can we extend previous knowledge by showing that the use of empirically supported faking indices as input outperforms the use of response patterns and scores?
7. Can we replicate the finding that logistic regressions are better than other approaches for detecting faking on IATs, whereas XGBoost is better for self-reports?

METHOD

Participants and Data Sets

- Set 1: *N* = 84 (28 faking low, 28 faking high, 28 control; 74 students; 64 women; average age: 22.37 years, *SD* = 4.45)
- Set 2: *N* = 197 (66 faking low, 67 faking high, 64 control; 196 students, 1 no response; 165 women; average age: 21.44 years, *SD* = 2.95)
- Set 3: *N* = 260 (88 faking low, 86 faking high, 86 control; 257 students; 191 women; average age: 21.22 years, *SD* = 4.74)
- Set 4: *N* = 293 (104 faking low, 94 faking high, 95 control; 293 students; 220 women; average age: 22.31 years, *SD* = 4.09)
- Set 5: *N* = 199 (67 faking low, 65 faking high, 67 control; 199 students; 163 women; average age: 21.53 years, *SD* = 3.18)
- Set 6: *N* = 299 (105 faking low, 97 faking high, 97 control; 299 students; 225 women; average age: 22.06 years, *SD* = 4.07)
- Set 7: *N* = 84 (28 faking low, 28 faking high, 28 control; 74 students; 64 women; average age: 22.37 years, *SD* = 4.45)

Procedure

- Participants worked on a baseline assessment and afterwards were randomly assigned to one of the following conditions: faking high scores, faking low scores, or working under the standard instructions of the measures (i.e., control condition)
- Whether they were asked to fake naïvely or whether they additionally received information about faking strategies varied between the studies (Table 1); also, whether they had faking practice varied between the studies (Table 1)
- ***Naïve Faking Without Faking Practice:*** assessed for four constructs: extraversion (Data Sets 1 to 4), conscientiousness (Data Set 5), need for cognition (Data Set 6), and self-esteem (Data Set 7)
- ***Naïve Faking With Faking Practice:*** assessed for two constructs: extraversion (Data Set 2) and conscientiousness (Data Set 5)
- ***Informed Faking Without Faking Practice:*** assessed for two constructs: extraversion (Data Set 1) and self-esteem (Data Set 7)
- ***Informed Faking With Faking Practice:*** assessed for self-esteem (Data Set 7)

Measures

Self-Reports

- *Extraversion Scale* from the NEO-Five Factor Inventory (Costa & McCrae, 1992)
- *Conscientiousness Scale* from the NEO-Five Factor Inventory (Costa & McCrae, 1992)
- *Need for Cognition Scale* (Cacioppo & Petty, 1982)
- *Rosenberg Self-Esteem Scale* (Rosenberg, 1965)

IATs

- *Extraversion IAT* (Back et al., 2009): self-relevant words (e.g., I, self) and non-self-relevant words (e.g., they, yours); extraversion-related words (e.g., talkative, active) and introversion-related words (e.g., shy, passive)
- *Conscientiousness IAT* (Steffens & Schulze-König, 2006): self-relevant words (e.g., I, self) and non-self-relevant words (e.g., they, yours); consciousness-related words (e.g., strong willed, pedantic) and nonconsciousness-related words (e.g., aimless, laid-back)
- *Need for Cognition IAT* (Fleischhauer et al., 2013): me words (e.g., me) and not me words (e.g., they, others); words related to reasoning (e.g., to scrutinize, to puzzle) and words related to relaxation (e.g., to chill out, to daydream)
- *Self-Esteem IAT* (Greenwald & Farnham, 2000; Rudolph et al., 2006): self-relevant words (e.g., I, self) and non-self-relevant words (e.g., they, yours); pleasant words (e.g., joy, smile) and unpleasant words (e.g., disaster, war)

Input Data

- ***Response Pattern*** (i.e., all IAT trials in IATS; all item responses in self-reports)
- ***Scores*** (i.e., D_2 and IAT_γ in IATs; test scores in self-reports)
- ***Faking Indices*** (i.e., CTS, IAT_α, IAT_γ, Ratio 150-10000, Slow_Co, IncErr_Co for the naïve faking and informed faking of low scores; CTS, IAT_α, Ratio 150-10000, Accel_Co for the naïve faking of high scores; CTS, IAT_α, IAT_{fg}, Ratio 150-10000, Slow_In for the informed faking of high scores (Cvencek et al., 2010; Röhner & Thoss, 2018; Agosta et al., 2011; Röhner et al., 2013)

Machine Learning, Multilayer Cross-Validation, and Performance Evaluation

- We used logistic regression, random forest, and XGBoost (Boldt et al., 2018; Calanna et al., 2020)
- We ran a five-fold cross-validation to tune the algorithms and additionally ran another 10-fold cross-validation to estimate their performance (see also Cawley & Talbot, 2010)
- We used the random search to find the best set of hyperparameters relative to the *F1* score in order to maximize the tradeoff between Precision and Recall (e.g., Calanna et al., 2020)

MAIN FINDINGS

Table 1

		Self-reports			IATs		
Data set	Measurement occasion	*M*	(*SD*)	Cronbach's α	*M*	(*SD*)	Split-half reliabilities
Extraversion							
1	Baseline	30.02	6.26	.75	0.24	0.44	.86
	Naive faking without practice	25.16	13.47	.94	0.16	0.57	.91
	Informed faking without practice	26.57	12.34	.95	0.19	0.86	.97
2	Baseline	29.77	6.27	.77	0.35	0.35	.73
	Naive faking without practice	26.41	13.68	.95	0.24	0.48	.81
	Naive faking with one practice trial	26.64	14.62	.96	0.20	0.47	.79
	Naive faking with two practice trials	27.06	14.77	.96	0.20	0.46	.70
	Naive faking with three practice trials	26.69	15.36	.97	0.20	0.47	.79
3	Baseline	28.00	6.26	.80	0.21	0.41	.84
	Naive faking without practice	26.03	14.34	.96	0.14	0.61	.88
4	Baseline	27.70	7.20	.85	0.12	0.43	.83
	Naive faking without practice	25.94	14.84	.97	0.13	0.56	.81
Conscientiousness							
5	Baseline	32.53	7.04	.86	0.56	0.30	.71
	Naive faking without practice	28.05	14.46	.97	0.45	0.43	.80
	Naive faking with one practice trial	27.73	15.65	.98	0.39	0.47	.80
	Naive faking with two practice trials	27.75	16.57	.98	0.37	0.44	.75
	Naive faking with three practice trials	27.39	16.59	.98	0.38	0.45	.79
Need for Cognition							
6	Baseline	16.02	11.88	.87	-0.04	0.44	.78
	Naive faking without practice	5.76	31.76	.98	0.00	0.57	.84
Self-Esteem							
7	Baseline	23.10	4.98	.87	0.70	0.28	.78
	Naive faking without practice	19.11	10.62	.98	0.47	0.48	.86
	Informed faking without practice	20.23	9.21	.97	0.30	0.93	.96
	Informed faking with one practice trial	20.01	9.54	.97	0.36	0.87	.93
	Informed faking with two practice trials	19.81	9.08	.97	0.37	0.78	.96

Note. Descriptives for self-reports were based on questionnaire data with a possible range from 0 to 4 (extraversion), 0 to 4 (conscientiousness), -3 to +3 (need for cognition), and 0 to 3 (self-esteem). Descriptives for the IAT were based on IAT data, which were treated with the recommended D_2 scoring algorithm (Greenwald et al., 2003a, 2003b). Split-half reliability was based on split-half correlations incorporating Spearman-Brown adjustments.

Faking Conditions Impacted Faking Detection with Machine Learning

- Naïve fakers were better detected on IATs than on self-reports; This was not true for informed faking
- When participants were naïve, faking detection was superior for extraversion and need for cognition than for conscientiousness and self-esteem; there were no differences when participants had practice in faking or were informed about how to fake
- Faking low was better detected than faking high
- Informed fakers were better detected than naïve ones
- Fakers could be detected comparably well at different practice levels

Input Data Impacted Faking Detection with Machine Learning

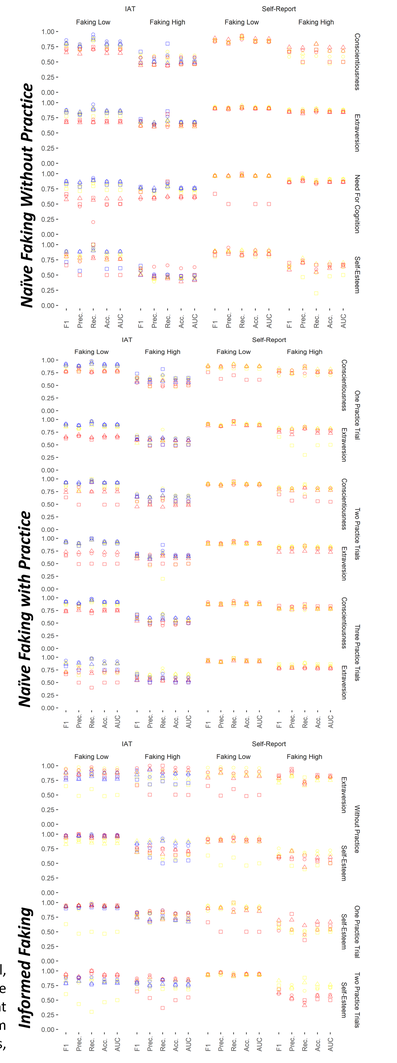
- Self-report measures: Response patterns and scores performed comparably well
- IAT: Faking indices and response patterns were superior to scores

Classifiers Impacted Faking Detection with Machine Learning

- Random forest and logistic regression were comparably good
- Both methods outperformed XGBoost

⇒ In most cases, classifiers were able to detect faking above chance
⇒ *F1* varied from .44 to .98

Performance Evaluations of Classifiers



Note. The five performance evaluation indices are presented on the x-axis. Prec. = Precision, Rec. = Recall, Acc. = Accuracy. Performance evaluation can range from 0.00 to 1.00 (y-axis). Geometrical shapes code the classifiers: Circles represent performance evaluations from logistic regression, triangles represent performance evaluations from random forest, and squares represent performance evaluations from XGBoost. Colors code the kind of input data: Yellow represents response patterns, red represents scores, and blue represents faking indices.

CONCLUSION AND IMPLICATIONS

Our research supports the assumption that there are different faking processes and showcases how machine learning has the potential to outperform human efforts in detecting faking. It also demonstrates that the circumstances that affect faking have to be taken into account.

ACKNOWLEDGMENTS

This research was partly funded by a grant from the equal opportunities office at the University of Bamberg. The funding source had no involvement in the study design or analyses.

MAIN REFERENCES AND CONTACT INFORMATION

Boldt, B., J., While, Z., & Breimer, E. (2018). Detecting compromised Implicit Association Test results using supervised learning, 17th IEEE International Conference on Machine Learning and Applications (ICMLA)., Orlando FL, 2018.
Calanna, .P, Lauriola, M., Saggino, A., Tommasi, M., Furlan, S. (2020). Using a supervised machine learning algorithm for detecting faking good in a personality self-report. Int J Select Assess., 28, 176–185. https://doi.org/10.1111/ijsa.12279.
Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. Journal of Research in Personality, 47, 330-338. doi:10.1016/j.jrp.2013.02.009
Röhner, J., & Thoss, P. (2018). EZ: An easy way to conduct a more fine-grained analysis of faked and nonfaked Implicit Association Test (IAT) data. The Quantitative Methods for Psychology, 14(1), 17–37. https://doi.org/10.20982/tqmp.14.1.p017

Contact: Dr. Jessica Röhner, Department of Psychology, University of Bamberg, Germany, E-Mail: jessica.roehner@uni-bamberg.de