Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Review

Psychometric properties of personality assessment using machine learning

Antonis Koutsoumpis

Technological advancements have enabled personality psychologists to move beyond traditional questionnaire-based assessment toward machine learning-based personality assessment (ML-PA). This manuscript provides a non-systematic overview of the validity and reliability of ML-PA, where behavioral features (e.g., text, voice, digital footprints) serve as predictors of personality traits. ML-PA shows promising construct validity, particularly for observer reports, and ML-PA values are similarly correlated with external variables as questionnaire-based values. However, reliability indices, especially for self-report-based ML-PAs, have been found to be lower. Factors such as sample size, input data quantity, and trait activation significantly impact ML-PA accuracy. Algorithmic bias might pose a threat to ML-PA, and there is a trade-off between applying bias mitigation techniques and maximizing ML-PA performance. Future advancements, including the use of large language models and a focus on explainability, are expected to further enhance personality measurement using computational methods.

Address
Department of Social Psychology, Tilburg School of Social and Behavioral Sciences, Tilburg University, Netherlands

Corresponding author: Koutsoumpis, Antonis (a.koutsoumpis@tilburguniversity.edu)

Introduction

Personality assessment has been traditionally based on questionnaires completed either by the self or others [1]. However, important technological developments introduced in the first decades of the 21st century, such as the increase of computational power, the expanded computer storage and memory capacities, as well as the accessibility to high-performing computing systems, have allowed personality psychologists and computer scientists to use complex computational models for personality assessment, such as (un)supervised machine learning and large language models.

Personality assessment using supervised machine learning models, which will be the focus of the present manuscript, has some important differences compared to the traditional, questionnaire-based personality research. In personality research, personality traits serve as the *independent* variables, and outcome variables, such as behaviors, serve as the dependent variables (Fig. 1a). Instead, in machine learning-based personality assessment (ML-PA; Fig. 1b), behavior becomes the independent variable and personality traits the *dependent* variables [2,31]. For example, a large number of behavioral features, such as text, voice characteristics, facial expressions, or digital footprints (e.g., Facebook 'Likes') — typically automatically collected using appropriate behavioral extraction software [3,4] — serve as the input (i.e., independent variables) to supervised machine learning models to predict questionnaire-based personality ratings (i.e., dependent variables) [5].

As can be seen in Fig. 1, ML-PA can only capture a proportion of questionnaire-based personality variance, but almost never the full variance. Consequently, measuring personality traits using machine learning scales (instead of questionnaire-based scales) to predict outcome variables might seem as a suboptimal approach. However, in many instances behavioral cues may be more readily available than questionnaire data. For instance, in a selection and assessment context, job candidates may already provide a large number of behavioral features, such as verbal and non-verbal behaviors during the job interview, or textual information contained in curriculum vitae and motivation letters. To the extent that ML-PA scales have rigorous psychometric properties, behavioral features may serve as a readily available and cost-effective substitute for questionnaire-based personality traits, which — in turn — can be used to predict future behavior. For instance, ML-PA scores derived from a one-way online video interview may be used to predict future job performance [6].

The present manuscript is a non-systematic overview of the validity and reliability of ML-PA (the area captured by $\widehat{P}$ in Fig. 1). For other thorough reviews on the topic,

Current Opinion in Psychology 2025, **65:**102093

This review comes from a themed issue on Personality

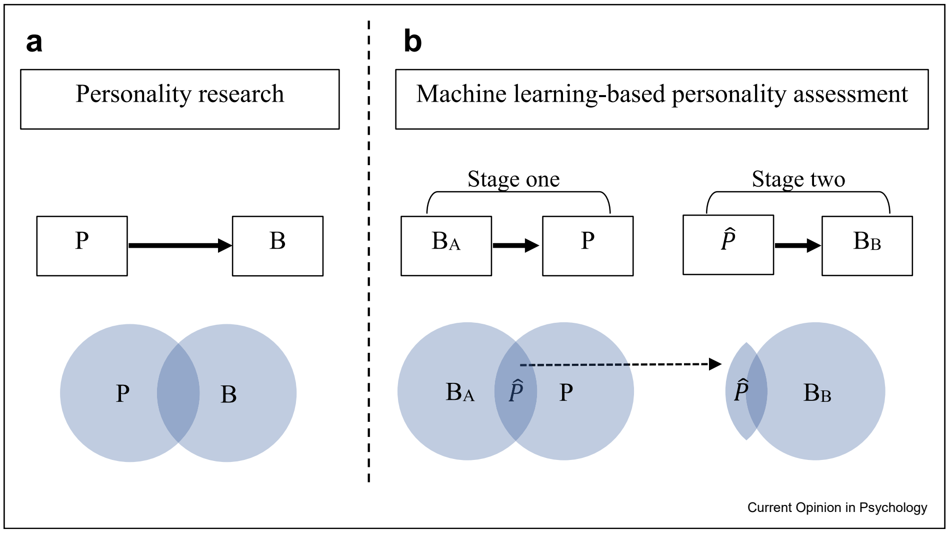
Edited by Reinout E. de Vries, Kibeom Lee and Michael C. Ashton

For complete overview about the section, refer Personality

Available online 28 June 2025

https://doi.org/10.1016/j.copsyc.2025.102093

Fig. 1


a
Personality research
P
B
P
B
b
Machine learning-based personality assessment
Stage one
Stage two
BA
P
P̂
BB
BA
P̂
P
P̂
BB
Current Opinion in Psychology

Questionnaire-based **(a)** and machine learning-based **(b)** personality assessment. *Notes.* P = personality traits (e.g., self-reports); B = behavior (e.g., job performance); B_A = behavior in context A (e.g., verbal behavior in job interview); $\hat{P}$ = predicted personality traits from behavioral features; B_B = behavior in context B (e.g., job performance); arrows with a continuous solid line refer to 'prediction'.

the readers are encouraged to also consult the following articles [2,5,7]. The next pages will provide a brief summary of the psychometric properties of ML-PA, as well as a discussion on algorithmic bias.

Validity

Construct validity

The performance of machine learning models can be evaluated using various effect sizes (e.g., Pearson *r*, R^2, Mean Squared Error, Accuracy, *F*-measure) [8,31]. In the present manuscript, the construct validity will be operationalized using the Pearson *r* correlation coefficient, calculated as the correlation between the ground truth (questionnaire-based) self- and/or observer reports of personality with the predicted (machine learning-based) self- and/or observer reports. In the remaining of the manuscript, this correlation will be referred to as the 'accuracy' of machine learning models.

Table 1 summarizes studies that have used supervised machine learning techniques to assess personality using self- and observer reports. Regarding self-reports, Hinds and Joinson [5] have meta-analytically summarized ML-PA accuracy, and Table 1 contains a few additional studies published after this meta-analysis, as well as studies that have collected observer reports. The average sample size weighted accuracy of ML-PA was $r = 0.30$ and 0.55, for self- and observer reports respectively, averaged across all personality traits.

Regarding self-reports, the effect sizes varied to a relatively small extent, ranging from $r = 0.27$ (Emotionality/Neuroticism) to $r = 0.34$ (Openness to Experience), showing that self-reported personality traits can be assessed somewhat to a similar extent. The effect size was somewhat smaller for Honesty-Humility ($r = 0.17$), but it was based on a single study. Regarding observer reports, the effect sizes varied to a greater extent, ranging from $r = 0.41$ (Agreeableness) to $r = 0.60$ (Extraversion; again, Honesty-Humility was somewhat lower; $r = 0.29$). These results are in line with previous literature suggesting that some personality traits, such as Extraversion, are more 'visible' than others [9,10].

The effect sizes in Table 1 also show that the accuracy of ML-PA was larger for observer compared to self-reported personality traits. This asymmetry (i.e., the fact that verbal and non-verbal behaviors were more strongly correlated with observer instead of self-reports of personality) is a prevalent finding in the literature [11,9]. A conceptual interpretation of the asymmetry is provided by the self-other knowledge asymmetry (SOKA) model [10], according to which observers have more access to evaluative traits and visible behaviors, such as someone's facial expressions, voice, or spoken text. Instead, the self has more access to non-evaluative traits and internal behaviors, such as their thoughts or feelings. A methodological interpretation is provided by the level of contextualization of personality measurement. Observer reports are based on the available verbal and non-verbal behaviors in a specific task (e.g., video interview, digital footprints). However, self-reports are based on questionnaire items the content of which is not necessarily related to the verbal and non-verbal

Table 1

Pearson *r* correlation coefficients between human-based and machine learning-predicted self- and observer reports of personality.

Study	*N*	Task	Input	H	E	X	A	C	O	*Mean*
Self-reports										
[5]	687,414[a]	Meta-analysis (various tasks)	Digital data		0.27	0.30	0.28	0.31	0.34	*0.30*
[17]	685	AVI	Audio, visual, verbal	0.17	0.43	0.42	0.08	0.25	0.26	*0.29*
[20]	1037	Chatbot	Verbal		0.55	0.59	0.57	0.53	0.63	*0.59*
[20]	407	Chatbot	Verbal		0.40	0.45	0.42	0.46	0.58	*0.48*
[23]	7691	CV and short text	Verbal		0.23	0.33	0.29	0.18	0.21	*0.25*
[24]	2045	Audio recording	Audio, verbal		0.39	0.26	0.38	0.36	0.29	*0.34*
Sample size weighted mean				**0.17**	**0.27**	**0.30**	**0.28**	**0.31**	**0.34**	***0.30***
Observer reports										
[17]	685	AVI	Audio, visual, verbal	0.29	0.37	0.63	0.33	0.49	0.40	*0.44*
[21]	1073	AVI	Audio, visual, verbal		0.27	0.65	0.34	0.42	0.35	*0.41*
[25]	408	YouTube	Audio, visual, verbal		0.48	0.69	0.62	0.47	0.41	*0.54*
[26]	36	Video interview	Audio, visual, verbal		0.40	0.44	0.38	0.34	0.35	*0.38*
[27]	10,000	YouTube (15 s clips)	Audio, visual, verbal		0.64	0.67	0.48	0.61	0.62	*0.60*
[28]	207	CEO earning calls	Verbal		0.62	0.65	0.67	0.64	0.67	*0.65*
[29]	939	YouTube	Audio, visual		0.00	0.53	0.22	0.17	0.45	*0.27*
Sample size weighted mean				**0.29**	**0.55**	**0.65**	**0.45**	**0.55**	**0.57**	**0.55**

Notes. N = sample size; H = Honesty-Humility; E = Emotionality (in HEXACO), Emotional Stability or Neuroticism (in the Big Five or Five Factor Models); X = Extraversion; A = Agreeableness (note that Agreeableness is operationalized and structured somewhat differently in the HEXACO and Big Five or Five Factor Models [43]); C = Conscientiousness; O = Openness to Experience; CV = Curriculum Vitae; AVI = Asynchronous Video Interview; CEO = Chief Executive Officer; the meta-analysis of [5] reported *corrected* Pearson *r* correlation coefficients.

[a] The sample size contained some dependency, mainly due to the MyPersonality dataset having been analyzed by multiple researchers who have used different sets of input features as independent variables (e.g., digital footprints, Facebook post updates).

behaviors that participants exhibited in a specific task (e.g., video interview, digital footprints) [12,45]. When the SOKA model and the level of measurement contextualization have been directly contrasted, the level of measurement contextualization seemed to mainly account for the asymmetry [13].

The accuracy of ML-PA also depends on the sample size and the amount of input data. Two studies have explored the accuracy of text-based ML-PA tools in predicting personality. Regarding self-reports, Eichstaedt et al. [18] found that, as sample size increased from zero up to n = 65,896, the accuracy of ML-PA increased following an inverse exponential function that reached an asymptote at approximately n = 5000. Regarding observer reports, Hickman et al. [19], analyzing a dataset of n = 1034, found that the accuracy of their ML-PA increased following an inverse exponential function that reached an asymptote at approximately n = 500 (thus, sooner than self-reports). In both abovementioned studies, the results were affected by the text analysis technique as well as by text length, with more advanced natural language processing models (e.g., Bidirectional Encoder Representations from Transformers [BERT] models vs. binary unigrams) reaching the asymptote sooner (e.g., n = 500 vs. n = 750; [19]), whereas longer texts (in combination with sample size) were also associated with higher accuracy [18]. The modeling process, such as the type of cross-validation, did not seem to affect the accuracy of ML-PA, at least for self-reports [5].

Finally, to maximize the accuracy of ML-PA, it is beneficial for the behavioral tasks to be aligned with the evaluated personality traits, as suggested by Trait Activation Theory (TAT) [14,15]. That is, if someone wants to measure a specific personality trait, participants should engage in a task that allows for the activation and behavioral expression of that specific personality trait. Indeed, TAT has received empirical support in multiple contexts [16], including ML-PA [17].

Nomological network, discriminant and convergent validity

The ML-PA scores have also been studied in relation to external variables, typically referred to as the nomological network or discriminant and convergent validity of ML-PA [7,2]. The desirable outcome would be for ML-PA values to be similarly correlated with external variables as do questionnaire-based personality scores. Table 2 presents the correlations between questionnaire- and machine learning-based personality traits with a variety of external variables. Regarding self-reports, the results show that ML-PA correlated somewhat more strongly with external variables compared to questionnaire-based self-reports (r = 0.08 and 0.13, for questionnaire and ML-PA), whereas for observer reports the difference was less pronounced (r = 0.16 and 0.15, for questionnaire and ML-

Table 2

Pearson *r* correlations between questionnaire- and machine learning-based personality traits with external variables.

Study	N	Construct	H		E		X		A		C		O		|Mean|	
			Q	ML	Q	ML	Q	ML	Q	ML	Q	ML	Q	ML	Q	ML
Self-reports																
[17]	685	Interview performance	0.03	0.07	–0.02	0.11	0.32	0.36	–0.02	–0.03	0.21	0.26	0.10	0.11	*0.12*	*0.16*
[17]	685	Physical attractiveness	–0.08	–0.07	0.00	0.15	0.12	0.13	0.01	–0.05	0.05	0.07	–0.04	0.00	*0.05*	*0.08*
[20]	289	College adjustment			0.16	0.13	0.08	0.18	0.14	0.12	0.25	0.17	–0.01	–0.02	*0.13*	*0.12*
[20]	379	GPA			–0.02	0.15	–0.13	0.08	0.14	0.04	0.33	0.14	–0.14	–0.03	*0.15*	*0.09*
[20]	407	ACT			0.00	0.08	–0.13	–0.13	0.05	–0.03	0.05	0.04	0.08	0.19	*0.06*	*0.09*
[23]	3469	VI: Design			–0.06	–0.16	–0.01	–0.11	0.03	–0.04	–0.07	–0.22	0.15	0.13	*0.06*	*0.13*
[23]	3469	VI: Marketing			–0.10	–0.19	0.11	0.19	0.06	0.12	–0.06	–0.07	0.06	0.07	*0.08*	*0.13*
[23]	3469	VI: Programming			0.04	0.05	–0.03	–0.23	–0.01	–0.13	–0.05	–0.13	0.08	0.09	*0.04*	*0.13*
[23]	3469	VI: Finance			0.13	0.42	0.08	0.26	0.03	0.22	0.10	0.33	–0.01	0.01	*0.07*	*0.25*
[23]	3469	VI: Analytics			0.10	0.28	0.03	0.03	–0.02	0.05	0.03	0.18	0.01	0.06	*0.04*	*0.12*
[23]	3469	VI: Operations			0.06	0.13	0.09	0.20	0.05	0.15	0.03	0.22	0.05	0.07	*0.06*	*0.15*
[23]	3469	VI: Accounting			0.02	0.13	0.01	0.10	–0.01	0.06	0.10	–0.01	–0.06	–0.13	*0.04*	*0.09*
[23]	3469	VI: Human resources			–0.06	–0.11	0.07	0.13	0.07	0.11	0.00	0.16	0.00	–0.08	*0.04*	*0.12*
[30]	1082	Satisfaction with life			0.46	0.19	0.24	0.13	0.24	0.21	0.29	0.19	0.05	–0.03	*0.26*	*0.15*
[30]	927	Self-monitoring			0.10	0.05	0.36	0.15	–0.03	–0.01	–0.03	–0.09	0.18	0.08	*0.14*	*0.08*
[30]	864	Fair mindedness			0.35	0.19	0.24	0.10	0.28	0.17	0.33	0.23	0.17	0.03	*0.27*	*0.14*
[30]	864	Self-disclosure			0.28	0.16	0.15	0.14	0.37	0.28	0.37	0.29	–0.02	–0.07	*0.24*	*0.19*
[30]	736	Recent physician visits			–0.14	–0.08	0.05	0.10	0.02	0.03	–0.05	0.12	0.00	–0.01	*0.05*	*0.07*
[30]	733	Recent days sick			–0.22	–0.11	–0.01	0.03	–0.02	0.02	–0.07	–0.01	0.01	0.07	*0.07*	*0.05*
[30]	549	Impulsiveness scale			–0.23	–0.07	0.10	0.12	–0.13	–0.15	–0.43	–0.10	0.00	0.01	*0.18*	*0.09*
[30]	1842	Facebook friends			0.12	0.09	0.18	0.23	0.05	0.07	0.05	0.10	0.04	0.00	*0.09*	*0.10*
[30]	756	Politically liberal			–0.05	–0.08	0.07	0.03	–0.01	–0.19	–0.13	–0.14	0.32	0.22	*0.12*	*0.13*
|*Sample size weighted mean*|			**0.06**	**0.07**	**0.10**	**0.17**	**0.09**	**0.15**	**0.06**	**0.11**	**0.09**	**0.16**	**0.06**	**0.07**	***0.08***	***0.13***
Observer reports																
[6]	107	Education			0.02	–0.01	–0.01	0.04	–0.09	–0.23	0.05	0.06	0.06	0.02	*0.05*	*0.07*
[6]	107	Years work experience			–0.03	0.19	0.11	0.17	–0.12	0.00	0.09	–0.05	–0.13	–0.14	*0.10*	*0.11*
[6]	107	Number of jobs held			–0.04	–0.21	0.08	–0.01	–0.21	–0.05	–0.05	–0.14	–0.01	–0.10	*0.08*	*0.10*
[6]	107	Hireability			0.60	0.26	0.67	0.54	0.68	0.56	0.73	0.51	0.62	0.47	*0.66*	*0.47*
[17]	685	Interview performance	0.22	0.31	0.07	0.09	0.59	0.56	0.24	0.36	0.41	0.50	0.37	0.44	*0.32*	*0.38*
[17]	685	Physical attractiveness	0.11	0.14	–0.08	0.02	0.25	0.14	0.15	0.19	0.16	0.16	0.17	0.16	*0.15*	*0.14*
[21]	161[a]	College GPA			–0.13	0.01	–0.04	–0.03	–0.03	0.04	0.15	0.08	–0.09	0.01	*0.09*	*0.03*
[21]	375[a]	High school GPA			0.04	0.08	0.15	0.22	0.15	0.13	0.09	0.16	0.10	0.12	*0.11*	*0.14*
[21]	302[a]	SAT verbal			–0.02	0.02	0.04	–0.01	–0.04	–0.09	0.17	0.11	0.15	–0.01	*0.08*	*0.04*
[21]	306[a]	SAT math			–0.04	0.09	–0.08	–0.16	–0.24	–0.23	0.24	0.01	0.10	–0.07	*0.14*	*0.11*
[21]	222[a]	ACT			0.10	0.13	0.09	0.15	–0.05	–0.08	0.17	0.15	0.29	0.12	*0.14*	*0.12*
[21]	270[b]	College GPA			0.06	0.01	0.09	0.01	0.21	0.08	0.19	0.00	0.17	–0.01	*0.14*	*0.02*
[21]	282[b]	High school GPA			0.03	0.00	0.06	0.13	0.12	0.11	0.10	0.10	0.05	0.10	*0.07*	*0.09*

[21]	225[b]	SAT verbal			0.13	0.01	0.10	0.12	0.03	−0.08	0.19	0.19	0.01	0.10	*0.09*	*0.10*
[21]	226[b]	SAT math			−0.04	−0.06	−0.06	−0.03	−0.16	−0.19	0.08	−0.03	−0.03	0.01	*0.07*	*0.06*
[21]	162[b]	ACT			0.21	0.07	0.10	0.11	0.12	−0.05	0.23	0.20	0.10	0.14	*0.15*	*0.11*
|Sample size weighted mean|			**0.17**	**0.23**	**0.08**	**0.06**	**0.20**	**0.19**	**0.16**	**0.17**	**0.20**	**0.18**	**0.17**	**0.15**	**0.16**	**0.15**

Notes. The 'Mean' and 'Sample size weighted mean' are calculated in absolute values; *N* = sample size; H = Honesty-Humility; E = Emotionality (in HEXACO), Emotional Stability or reverse Neuroticism (in the Big Five or Five Factor Models); X = Extraversion; A = Agreeableness (note that Agreeableness is operationalized and structured somewhat differently in the HEXACO and Big Five or Five Factor Models [43]); C = Conscientiousness; O = Openness to Experience; Q = Questionnaire-based personality scores; ML = Machine learning-based personality scores; GPA = Grade Point Average; ACT = American College Testing; VI = Vocational Interests; in Ref. [21] the correlation coefficients for Samples 2 and 3 have been averaged from the values in Ref. [21]'s Tables 9 and 11, respectively.

[a] Sample 2.

[b] Sample 3.

PA). At closer inspection, it seemed that for self-reports, questionnaire-based scores correlated more strongly with self-reported external variables, such as Satisfaction with Life or Self-monitoring, whereas ML-PA scores correlated more strongly with observer reported external variables, such as interview performance or attractiveness (possibly an indication of shared method variance).

However, averaging across multiple personality traits and external variables might mask underlying differences. A more suitable measure might be to test whether the correlations between questionnaire and ML-PA-based values with external variables point towards the same direction (e.g., an external variable is positively/negatively correlated both with questionnaire and ML-PA personality scores). When focusing only on the Pearson r values that were statistically significant both for questionnaire and ML-PA values (significance not shown in Table 2), all correlations (100 %) with external variables were in the same direction. Alternatively, one could calculate the profile correlation between questionnaire and ML-PA effect sizes (e.g., in Table 2, the correlation between the columns 'questionnaire-based Extraversion' and 'ML-PA-based Extraversion'). When focusing only on the Pearson r values that were statistically significant both for questionnaire and ML-PA values, the profile correlations (applying the r-to-z-to-r Fisher transformation) between questionnaire and ML-PA were on average $r = 0.64$ and 0.73, for self- and observer reports, respectively. Thus, it seems that questionnaire and ML-PA values show a substantial agreement in their correlations with external variables.

The convergent and discriminant validity of ML-PA has also been calculated using a multi-trait multi-method matrix, taking into account the different personality traits (i.e., multi-trait) and the two ways of assessing those personality traits, namely questionnaire and ML-PA (i.e., multi-method) [20,21]. Such results typically show good convergent and discriminant validity for observer reports, and somewhat weaker convergent and discriminant validity for self-reports.

Criterion validity

Criterion validity has been scarcely measured in ML-PA. In one study [6], ML-PA from asynchronous video interviews were non-significantly correlated with future job performance. However, the results were based on a small sample of $n = 25$ that did not allow to draw meaningful conclusions. Criterion validity is an area of research for which more empirical findings are needed to fully understand the usefulness of ML-PA.

Reliability

A small number of studies has explored the reliability of machine learning personality assessment (Table 3). Reliability has been measured in terms of (a) Cronbach's alpha (average $\alpha = 0.88$), (b) split-half reliability

Table 3

Reliability indices of machine-learning personality assessment.

Study	*N*	Reliability index	H	E	X	A	C	O	*Mean*
Self-reports									
[20]	1037	Cronbach’s a		0.90	0.90	0.92	0.92	0.76	*0.88*
[20]	407	Cronbach’s a		0.88	0.90	0.91	0.93	0.76	*0.88*
[21]	99	GCES		0.29	0.02	0.04	0.09	−0.03	*0.08*
[20]	1037	Split-half reliability		0.60	0.64	0.73	0.67	0.68	*0.66*
[20]	407	Split-half reliability		0.57	0.64	0.68	0.68	0.63	*0.64*
[17]	145	Test-retest	0.30	0.64	0.32	0.41	0.33	0.27	*0.39*
[20]	61	Test-retest		0.58	0.66	0.63	0.59	0.67	*0.63*
[21]	99	Test-retest		0.76	0.34	0.56	0.30	0.32	*0.46*
[30]	331−1424	Test-retest		0.62	0.72	0.65	0.76	0.74	*0.70*
Observer reports									
[21]	99	GCES		0.05	0.66	0.19	0.53	0.31	*0.35*
[17]	145	Test-retest	0.46	0.21	0.58	0.26	0.54	0.44	*0.41*
[21]	99	Test-retest		0.23	0.70	0.43	0.65	0.52	*0.51*
[28]	1563	Test-retest		0.85	0.75	0.83	0.81	0.83	*0.81*

Notes. N = sample size; H = Honesty-Humility; E = Emotionality (in HEXACO), Emotional Stability or Neuroticism (in the Big Five or Five Factor Models); X = Extraversion; A = Agreeableness (note that Agreeableness is operationalized and structured somewhat differently in the HEXACO and Big Five or Five Factor Models [43]); C = Conscientiousness; O = Openness to Experience; GCES = Generalized coefficient of equivalence and stability.

Table 4

Presence and absence of bias in ground truth measures and the modeling process.

		Ground truth	
		No bias	Bias
Modeling process	No bias	A No ground truth bias No modeling bias	B Ground truth bias No modeling bias
	Bias	C No ground truth bias Modeling bias	D Ground truth bias Modeling bias

Note. ‘Ground truth’ refers to the original (human-based) data, used to train the machine learning model (e.g., questionnaire based self-reports of personality).

(average $r_{split\text{-}half}$ = 0.65), (c) test-retest reliability (sample size weighted average $r_{test\text{-}retest}$ = 0.63 and 0.76 for self- and observer reports, respectively), as well as (d) the generalized coefficient of equivalence and stability (GCES; *GCES* = 0.08 and 0.35, for self- and observer reports, respectively). The GCES [44] measures whether an algorithm trained on one dataset generalizes to other datasets on which it was not trained. As such, it is more a measure of generalizability across datasets instead of reliability within a single occasion or task (e.g., Cronbach’s alpha), and, as a result, it shows lower values compared to other reliability indices. From a statistical standpoint, test-retest reliability seems to be the preferred measure of reliability, as it examines the degree to which the ML-PA scores from one instance generalize to another similar instance [2], whereas the GCES seems to be the preferred measure of generalizability [44].

Algorithmic bias

Algorithmic bias (e.g., ML-PA bias) is a sensitive topic with possible negative consequences, especially when algorithms are used to make automated decisions [32−35]. There are two broad sources of bias in machine learning: (a) bias in ground truth scores (e.g., more attractive job applicants receiving higher job interview scores [36]), and (b) bias in the modeling process (e.g., non-representative data split in cross-validation) [32,33,35,38]. Note, however, that sometimes subgroup differences in ground truth scores, such as gender differences in Emotionality [39], represent factual differences and do not constitute instances of bias. Table 4 provides an overview of machine learning bias instances taking into account the presence or absence of bias in the ground truth and modeling process. Quadrant A describes the absence of bias. Quadrants B-D describe instances of bias either when only ground truth (B) or modeling process (C) bias are present, or a combination of both (D). Currently, there does not seem to be a consensus regarding the definition of algorithmic bias in the literature, for example whether quadrant B should be labeled as ‘algorithmic bias’, as quadrants C-D do [37].

Depending on the source of bias, researchers can apply different bias mitigation techniques [32,40]. However, there is a trade-off between maximizing performance and mitigating possible biases, as bias mitigation techniques frequently reduce the overall performance of ML-PA models [38,41,42]. Algorithmic bias has been studied in multiple applied settings [32–35], but scarcely in relation to personality traits (e.g. [17]), where more research is needed.

Conclusion

In the last few decades, a substantial number of studies have employed machine learning models in assessing personality traits with the results showing promising validity and somewhat less optimal reliability indices. Observer reports of personality are more accurately evaluated compared to self-reports, and the accuracy of ML-PA improves when using large amounts of input features and observations, as well as with proper activation of personality traits. Notably, even though machine learning models are powerful in predicting personality traits, large language models (LLM) are expected to play a crucial role in the coming years, especially as LLMs show somewhat comparable accuracies to ML-PA, even when the latter has been carefully designed for personality assessment [22]. Finally, personality measurement will benefit from the introduction of new computational methods especially if the new techniques focus also on explainability (on top of predictive accuracy), as well as on addressing possible instances of algorithmic bias.

Credit author statement

Antonis Koutsoumpis: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Software, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration.

Declaration of competing interest

The author declares no conflict of interest.

References

References of particular interest have been highlighted as:

* of special interest
** of outstanding interest

1. Ashton MC, Lee K: **Self-and observer reports of personality**. *Annu Rev Psychol* 2025, **76**:771–795, https://doi.org/10.1146/annurev-psych-020124-115044.

2. Stachl C, Pargent F, Hilbert S, *et al.*: **Personality research and assessment in the era of machine learning**. *Eur J Pers* 2020, **34**:613–631, https://doi.org/10.1002/per.2257.

3. Baltrusaitis T, Zadeh A, Lim YC, Morency LP: **Openface 2.0: facial behavior analysis toolkit**. In *2018 13th IEEE int conf autom face gesture recognit*; 2018:59–66, https://doi.org/10.1109/FG.2018.00019.

4. Boersma P, Weenink D: *Praat: doing phonetics by computer [computer program]. Version 6.3.08*. 2023. http://www.praat.org/. Accessed 1 June 2025.

5. ** Hinds J, Joinson AN: **Digital data and personality: a systematic review and meta-analysis of human perception and computer prediction**. *Psychol Bull* 2024, **150**:727–766, https://doi.org/10.1037/bul0000430.

6. Stevenor BA, Hickman L, Zickar MJ, Wimbush F, Beck W: **Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews**. *Int J Sel Assess* 2024, **32**:544–560, https://doi.org/10.1111/ijsa.12480.

7. Bleidorn W, Hopwood CJ: **Using machine learning to advance personality assessment and theory**. *Pers Soc Psychol Rev* 2019, **23**:190–203, https://doi.org/10.1177/1088868318772990.

8. Phan LV, Rauthmann JF: **Personality computing: new frontiers in personality assessment**. *Soc Personal Psychol Compass* 2021, **15**, e12624, https://doi.org/10.1111/spc3.12624.

9. Koutsoumpis A, Oostrom JK, Holtrop DJ, Van Breda W, Ghassemi S, De Vries RE: **The kernel of truth in text-based personality assessment: a meta-analysis of the relations between the big five and the Linguistic inquiry and word count (LIWC)**. *Psychol Bull* 2022, **148**:843–868, https://doi.org/10.1037/bul0000381.

10. Vazire S: **Who knows what about a person? The self-other knowledge asymmetry (SOKA) model**. *J Pers Soc Psychol* 2010, **98**:281–300, https://doi.org/10.1037/a0017908.

11. Breil SM, Osterholz S, Nestler S, Back MD: **Contributions of nonverbal cues to the accurate judgment of personality traits**. In *The Oxford handbook of accurate personality judgment*. Edited by Letzring TD, Spain JS, Oxford University Press; 2021: 195–218, https://doi.org/10.1093/oxfordhb/9780190912529.013.13.

12. Nisbett RE, Caputo C, Legant P, Marecek J: **Behavior as seen by the actor and as seen by the observer**. *J Pers Soc Psychol* 1973 Aug, **27**:154–164, https://doi.org/10.1037/h0034779.

13. Koutsoumpis A: *Multimodal personality assessment from audio, visual, and verbal features*. Doctoral dissertation. Vrije Universiteit Amsterdam; 2024, https://doi.org/10.5463/thesis.714.

14. Tett RP, Burnett DD: **A personality trait-based interactionist model of job performance**. *J Appl Psychol* 2003, **88**:500–517, https://doi.org/10.1037/0021-9010.88.3.500.

15. Tett RP, Guterman HA: **Situation trait relevance, trait expression, and cross-situational consistency: testing a principle of trait activation**. *J Res Pers* 2000, **34**:397–423, https://doi.org/10.1006/jrpe.2000.2292.

16. Tett RP, Toich MJ, Ozkum SB: **Trait activation theory: a review of the literature and applications to five lines of personality dynamics research**. *Annu Rev Organ Psychol Organ Behav* 2021, **8**:199–233, https://doi.org/10.1146/annurev-orgpsych-012420-062228.

17. * Koutsoumpis A, Ghassemi S, Oostrom JK, *et al.*: **Beyond traditional job interviews: psychometric analysis of asynchronous job video interviews for personality and interview performance evaluation using machine learning**. *Comput Hum Behav* 2024, **154**, 108128, https://doi.org/10.1016/j.chb.2023.108128.

18. Eichstaedt JC, Kern ML, Yaden DB, *et al.*: **Closed and open vocabulary approaches to text analysis: a review, quantitative comparison, and recommendations**. *Psychol Methods* 2021, **26**:398–427, https://doi.org/10.1037/met0000349.

19. Hickman L, Liff J, Rottman C, Calderwood C: **The effects of the training sample size, ground truth reliability, and NLP method on language-based automatic interview scores' psychometric properties**. *Organ Res Methods* 2024, https://doi.org/10.1177/10944281241264027.

20. * Fan J, Sun T, Liu J, *et al.*: **How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores**. *J Appl Psychol* 2023, **108**: 1277–1299, https://doi.org/10.1037/apl0001082.

21. Hickman L, Bosch N, Ng V, *et al.*: **Automated video interview personality assessments: reliability, validity, and generalizability investigations**. *J Appl Psychol* 2022, **107**:1323–1351, https://doi.org/10.1037/apl0000695.

22. Zhang T, Koutsoumpis A, Oostrom JK, *et al.*: **Can large language models assess personality from asynchronous video interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns**. *IEEE Trans Affect Comput* 2024, **15**:1769–1785, https://doi.org/10.1109/TAFFC.2024.3374875.

23. Grunenberg E, Peters H, Francis MJ, Back MD, Matz SC: **Machine learning in recruiting: predicting personality from CVs and short text responses**. *Front Soc Psychol* 2024, **1**, 1290295, https://doi.org/10.3389/frsps.2023.1290295.

24. Lukac M: **Speech-based personality prediction using deep learning with Acoustic and linguistic embeddings**. *Sci Rep* 2024, **14**:1–13, https://doi.org/10.1038/s41598-024-81047-0.

25. Biel JI, Tsiminaki V, Dines J, Gatica-Perez D: **Hi youtube!: personality impressions and verbal content in social video**. In *Proceedings of the 15th ACM on international conference on multimodal interaction*; 2013, https://doi.org/10.1145/2522848.2522894.

26. Chen L, Feng G, Leong CW, *et al.*: **Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm**. In *Proceedings of the 18th ACM international conference on multimodal interaction*; 2016:161–168, https://doi.org/10.1145/2993148.2993203.

27. Ghassemi S, Zhang T, Van Breda WRJ, *et al.*: **Unsupervised multimodal learning for dependency-free personality recognition**. *IEEE Trans Affect Comput* 2023, **15**:1053–1066, https://doi.org/10.1109/TAFFC.2023.3318367.

28. Harrison JS, Thurgood GR, Boivie S, Pfarrer MD: **Measuring CEO personality: developing, validating, and testing a linguistic tool**. *Strat Mgmt J* 2019, **40**:1316–1330, https://doi.org/10.1002/smj.3023.

29. Nguyen LS, Frauendorfer D, Mast MS, Gatica-Perez D: **Hire me: computational inference of hirability in employment interviews based on nonverbal behavior**. *IEEE Trans Multimed* 2014, **16**:1018–1031, https://doi.org/10.1109/TMM.2014.2307169.

30. Park G, Schwartz HA, Eichstaedt JC, *et al.*: **Automatic personality assessment through social media language**. *J Pers Soc Psychol* 2015, **108**:934–952, https://doi.org/10.1037/pspp0000020.

31. Putka DJ, Beatty AS, Reeder MC: **Modern prediction methods: new perspectives on a common problem**. *Organ Res Methods* 2018 Jul, **21**:689–732, https://doi.org/10.1177/1094428117697041.

32. * * Hickman L, Huynh C, Gass J, Booth B, Kuruzovich J, Tay L: **Whither bias goes, I will go: an integrative, systematic review of algorithmic bias mitigation**. *J Appl Psychol* 2024 Dec 30, https://doi.org/10.1037/apl0001255.

33. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A: **A survey on bias and fairness in machine learning**. *ACM Comput Surv* 2021 Jul 13, **54**:1–35, https://doi.org/10.1145/3457607.

34. Köchling A, Wehner MC: **Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development**. *Bus Res* 2020 Nov, **13**:795–848, https://doi.org/10.1007/s40685-020-00134-w.

35. Akter S, McCarthy G, Sajib S, Michael K, Dwivedi YK, D'Ambra J, Shen KN: **Algorithmic bias in data-driven innovation in the age of AI**. *Int J Inf Manag* 2021 Oct 1, **60**, 102387, https://doi.org/10.1016/j.ijinfomgt.2021.102387.

36. Hosoda M, Stone-Romero EF, Coats G: **The effects of physical attractiveness on job-related outcomes: a meta-analysis of experimental studies**. *Pers Psychol* 2003 Jun, **56**:431–462, https://doi.org/10.1111/j.1744-6570.2003.tb00157.x.

37. Danks D, London AJ: **Algorithmic bias in autonomous systems**. *In: IJCAI* 2017 Aug, **17**:4691–4697. https://dl.acm.org/doi/abs/10.5555/3171837.3171944.

38. Tay L, Woo SE, Hickman L, Booth BM, D'Mello S: **A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment**. *Adv Methods Pract Psychol Sci* 2022 Feb, **5**, https://doi.org/10.1177/25152459211061337. 25152459211061337.

39. Lee K, Ashton MC: **Sex differences in HEXACO personality characteristics across countries and ethnicities**. *J Pers* 2020 Dec, **88**:1075–1090, https://doi.org/10.1111/jopy.12551.

40. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilović A, Nagar S: **AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias**. *IBM J Res Dev* 2019 Sep 18, **63**:4, https://doi.org/10.1147/JRD.2019.2942287. 1.

41. Booth BM, Hickman L, Subburaj SK, Tay L, Woo SE, D'Mello SK: **Integrating psychometrics and computing perspectives on bias and fairness in affective computing: a case study of automated video interviews**. *IEEE Signal Process Mag* 2021 Oct 27, **38**:84–95, https://doi.org/10.1109/MSP.2021.3106615.

42. Zhang N, Wang M, Xu H, Koenig N, Hickman L, Kuruzovich J, Ng V, Arhin K, Wilson D, Song QC, Tang C: **Reducing subgroup differences in personnel selection through the application of machine learning**. *Pers Psychol* 2023 Dec, **76**:1125–1159, https://doi.org/10.1111/peps.12593.

43. Ashton MC, Lee K, De Vries RE: **The HEXACO honesty-humility, agreeableness, and emotionality factors: a review of research and theory**. *Pers Soc Psychol Rev* 2014 May, **18**: 139–152, https://doi.org/10.1177/1088868314523838.

44. Gnambs T: **Facets of measurement error for scores of the big five: three reliability generalizations**. *Pers Indiv Differ* 2015 Oct 1, **84**:84–89, https://doi.org/10.1016/j.paid.2014.08.019.

45. Colvin CR, Funder DC: **Predicting personality and behavior: a boundary on the acquaintanceship effect**. *J Pers Soc Psychol* 1991 Jun, **60**:884–894, https://doi.org/10.1037/0022-3514.60.6.884.

Further information on references of particular interest

5. * * The authors performed two meta-analyses on the relation between digital data (e.g., Facebook 'likes', GPS data, spoken text in videos) and personality traits. The first meta-analysis (k = 30, n = 24,124) calculated the self-other agreement, where 'others' assessed personality traits from digital data, showing an average self-other agreement of ρ = 0.46. The second multilevel meta-analysis (k = 42, effect sizes n = 534) calculated the correlation between self-reports and machine learning-predicted self-reports from digital data, showing an average correlation of ρ = 0.30.

17. * Testing the psychometric properties of a machine-learning model to evaluate the personality traits of Extraversion and Conscientiousness from an asynchronous video interview (n = 710). The results showed that the machine learning model explained a larger proportion of observer (R^2 = 0.32), rather than self- (R^2 = 0.12), reports of personality, and the explained variance was larger when the algorithm was trained on trait-activating interview questions. The test-retest reliability (7 months interval; n = 145) was relatively weak (r_{tt} = 0.33 and 0.56, for self- and observer reports), and there were only few instances of algorithmic bias.

20. * The authors examined the psychometric properties of the machine-inferred personality scores (n = 1444), including reliability (internal consistency, split-half, and test–retest), factorial validity, convergent and discriminant validity, and criterion-related validity. The results showed that machine-inferred personality scores had overall acceptable reliability, comparable factor structure to questionnaires, good convergent validity but relatively poor discriminant validity, low criterion-related validity, and (occasional) incremental validity over self-reports.

32. * * The authors present a four-stage model of developing machine learning assessments and applying bias mitigation methods across the cycle of machine learning models, including (a) generating the training data, (b) training the model, (c) testing the model, and (d) deploying the model. The paper also includes a systematic review (k = 328) of instances of algorithmic biases in the literature.