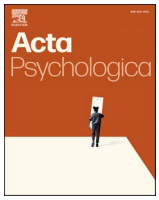
Acta Psychologica 228 (2022) 103659

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Scoring a forced-choice image-based assessment of personality: A comparison of machine learning, regression, and summative approaches

Airlie Hilliard [a,b,*], Emre Kazim [b,c], Theodoros Bitsakis [d], Franziska Leutner [a,d]

[a] Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[b] Holistic AI, London, UK
[c] Department of Computer Science, University College London, Gower St, London WC1E 6EA, UK
[d] HireVue, London, UK

ARTICLE INFO

Keywords:
Forced-choice
Image-based
Personality
Prediction
Machine learning
Psychometrics

ABSTRACT

Recent years have seen rapid advancements in the way that personality is measured, resulting in a number of innovative predictive measures being proposed, including using features extracted from videos and social media profiles. In the context of selection, game- and image-based assessments of personality are emerging, which can overcome issues like social desirability bias, lack of engagement and low response rates that are associated with traditional self-report measures. Forced-choice formats, where respondents are asked to rank responses, can also mitigate issues such as acquiescence and social desirability bias. Previously, we reported on the development of a gamified forced-choice image-based assessment of the Big Five personality traits created for use in selection, using Lasso regression for the scoring algorithms. In this study, we compare the machine-learning-based Lasso approach to ordinary least squares regression, as well as the summative approach that is typical of forced-choice formats. We find that the Lasso approach performs best in terms of generalisability and convergent validity, although the other methods have greater discriminate validity. We recommend the use of predictive Lasso regression models for scoring forced-choice image-based measures of personality over the other approaches. Potential further studies are suggested.

1. Introduction

In this article, we compare machine-learning-based, ordinary least squares, and summative approaches to scoring a forced-choice image-based assessment of personality, which we previously reported on the creation and validation of (Hilliard et al., 2022). While in recent years new ways of scoring forced-choice assessments have been developed that can overcome issues associated with traditional forced-choice scoring approaches (Brown & Maydeu-Olivares, 2011, 2013), these are typically for multidimensional measures. Since our measure has a combination of unidimensional and multidimensional items, these methods have limited applicability. As such, we previously used machine-learning-based scoring algorithms to overcome these challenges. Here we extend this work, examining how the use of different predictor combinations in different models impacts the validity of the measure. We begin by examining the significance of personality and how it is measured, both using traditional and more contemporary approaches, before narrowing our focus to image-based and forced-choice measures. We then describe the development of the models and evaluate their performance in terms of convergent and discriminant validity with the IPIP-NEO-120 and generalisability from the training to test data. In line with prior research (Speer & Delacruz, 2021), we conclude that machine-learning-based approaches outperform other scoring approaches and that they are a viable alternative option for scoring forced-choice assessments.

1.1. Measuring personality

An individual's personality has significant implications for many aspects of their life, including their wellbeing, social relationships, health, and career success (Roberts et al., 2007; Soldz & Vaillant, 1999). Indeed, the Big Five personality traits (openness to experience, conscientiousness, extraversion, agreeableness and neuroticism or emotional stability) are routinely tested in pre-employment screenings due to their ability to predict future job performance (Barrick & Mount, 1991; Kuncel et al., 2010; Pletzer et al., 2021; Rothmann & Coetzer, 2003;

* Corresponding author.
E-mail address: ahill015@gold.ac.uk (A. Hilliard).

https://doi.org/10.1016/j.actpsy.2022.103659
Received 9 March 2022; Received in revised form 22 June 2022; Accepted 22 June 2022
Available online 30 June 2022

Schmidt & Hunter, 1998; Schmitt, 2014). While self-report methods, such as the International Personality Item Pool (IPIP; Goldberg, 1992) scales and the NEO-PI R (Costa & McCrae, 2008), have been the default method of assessing personality until recently, self-report scales are associated with poor response quality (Krosnick, 1991) and incomplete responses due to respondent attrition (Yan et al., 2011), particularly if scales are lengthy. Self-reported measures of personality are also associated with social desirability bias or faking (van de Mortel, 2008), especially in high-stakes contexts, where respondents inflate their scores more compared to respondents completing the assessment in low stakes contexts (Arthur et al., 2010). This has implications for the use of personality assessments in high stakes contexts like recruitment, where candidates may attempt to inflate their scores to appear more favourably (Le et al., 2011).

1.2. Alternative ways of measuring personality

To overcome some of the issues associated with self-report measures, some have proposed using daily adjective-based measures of specific personality traits to avoid issues with one-time measurements (Di Sarno et al., 2020) while others have proposed more contemporary measures to predict personality from a wide range of sources. For example, personality has been predicted using the facial expressions of individuals in YouTube videos (Biel et al., 2012) and video interviews (Suen et al., 2019) and audio (speaking activity and prosody) and non-verbal cues (looking activity, pose and body movements) of vloggers in YouTube videos based on annotations of personality (Biel & Gatica-Perez, 2013). Others have inferred personality, based on observational ratings, from video resumes using features such as speaking activity, prosody, head motion, full face events, and overall motion (Nguyen & Gatica-Perez, 2016). Moving away from video analysis, personality has also been predicted from mobile phone data including calls and text frequency, GPS data and text response rate (de Montjoye et al., 2013), as well as through eye movement while running errands (Hoppe et al., 2018) and Facebook Likes (Kosinski et al., 2013). Others have used a text-based approach, using language to predict personality. While this is nothing new given that the Big Five model of personality was derived from language analysis (Digman, 1990), contemporary approaches use non-traditional sources of language such as social media posts instead of essays or descriptions of people and combine them with natural language processing computational techniques. For example, based on the frequency of word use and clusters of topics mentioned in Facebook status updates, personality has been predicted using latent Dirichlet allocation, a natural language processing technique used to cluster words into related topics (Park et al., 2015). Such approaches, therefore, move away from the need for self-report, reducing the influence of faking and allowing personality to be measured automatically (e.g. Park et al., 2015), although impression management on social media is not uncommon (Schlosser, 2020) so measures could still be affected by social desirability bias.

Something that is also gaining traction is image-based assessments of personality, which ask respondents to select images from a predefined set and use these choices to predict personality through machine-learning-based algorithms. Assessments of this format are being offered by commercial providers (e.g., HireVue and Traitify), and are also being investigated in the literature. For example, Leutner et al.' (2017) image-based measure of creativity presented respondents with sets of images and asked them to indicate which image was most like them. Using predictive algorithms, they predicted creativity scores on traditional scales, finding that the assessment could accurately measure curiosity ($r = 0.35$), cognitive flexibility ($r = 0.50$) and openness to experience ($r = 0.50$). Although this measure only assessed the openness to experience personality trait, more recent image-based measures have explored measuring all five traits through image choices. For example, Krainikovsky et al. (2019) presented respondents with 300 images tagged with information relating to the behavior, objects, emotions, and scenery in a set of images and asked them to select 20 to 100 preferred images. Based on the percentage of chosen pictures which related to specific tags, they predicted Big Five scores, with convergence with the NEO PI ranging from $r = 0.06$ for neuroticism to $r = 0.28$ for agreeableness. More recently, an image-based assessment of personality designed for use in selection used image choices from pairs of images mapped to the Big Five traits to predict personality, with convergence with the IPIP-NEO-120 ranging from 0.60 for agreeableness to 0.78 for extraversion (Hilliard et al., 2022).

The advantage of image-based assessments is that they are language neutral, meaning that they can remove language barriers to make the assessment more accessible to individuals like those with dyslexia (De Beer et al., 2014). They can also be taken by people speaking different languages without having to create language specific items (Paunonen et al., 1990), like is needed with questionnaire-based measures (Zhang et al., 2017). Images appear to elicit stronger preferences from respondents than questionnaire-based measures (Meissner & Rothermund, 2015), which could potentially contribute towards more rapid measurement if less time is spent deliberating. In addition, if the image-based measures are gamified (e.g., Hilliard et al., 2022) by adding features like sound effects and progress bars (Landers et al., 2021), this can have additional benefits since game-based assessments elicit less test-taking anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), are generally quicker to complete (Georgiou & Nikolaou, 2020; Leutner et al., 2020), and can be more engaging for respondents (Lieberoth, 2015). Assessments in this format can, therefore, overcome some of the issues associated with traditional self-report measures. This is particularly beneficial in the domain of recruitment since applicant perceptions of a selection process can influence how likely they are to accept a job offer (Hausknecht et al., 2004), which can have implications both for candidates and hiring managers who could potentially miss out on top talent if they have an unengaging recruitment process.

1.3. Forced-choice assessments

Forced-choice assessments vary from traditional self-report assessments in that they ask respondents to indicate the responses that are most and least like them, instead of asking them where they lie along a scale. Assessments of this type typically have two or four response options, with the most common format being multidimensional i.e. showing blocks of response options with statements from multiple constructs (Hontangas et al., 2015). In the context of personality, a respondent could be shown statements relating to multiple traits and asked to identify the statements they identify with most and least. Assessments of this format can prevent central tendency and extreme response styles (Brown & Maydeu-Olivares, 2013) since there is no midpoint, as well as acquiescence responding, where respondents select both positive and negative statements, since they are not able to endorse all of the statements presented to them (Brown & Maydeu-Olivares, 2013). Further, they are more resistant to faking than traditional measures (Salgado & Táuriz, 2014), which is particularly important in high-stakes contexts like recruitment.

The most typical way of scoring two-item forced-choice measures is by allocating a score of 1 when a positive item is selected (e.g., high extraversion) and a score of 0 when a negative item is selected (e.g., low extraversion). When the measure is multidimensional, featuring positive statements about two different traits, the selected statement is given a score of 1 and the unselected a score of 0. Therefore, the total score for that trait is calculated by summing the number of positive items selected relating to that trait (Hontangas et al., 2015). When the blocks have more than two statements and respondents are asked to indicate the most and least like them, the item selected for most is given 2 points, the unselected 1 point, and the least favoured 0 points. Again, scores for each construct are calculated by summing the number of points relating to each trait (Hontangas et al., 2015).

Due to the fact that multidimensional forced-choice measures result

in ipsative scales, where the score on one dimension is relative to another dimension and the overall score for each respondent across the constructs is the same, concerns have been raised about how well individuals can be compared since it is impossible to score above or below the mean score for all constructs (Brown & Maydeu-Olivares, 2011). However, mixed-dimensional forced-choice formats, or those combining multidimensional and unidimensional items, are not prone to ipsative scoring in the same way as fully multidimensional measures as they behave more like a traditional questionnaire-based method. To combat the issue of ipsative scores with multidimensional measures, alternative ways of scoring forced-choice measures have been proposed, based on Item Response Theory (IRT) and Thurstone's framework for comparative data (Brown & Maydeu-Olivares, 2011, 2013). These models, which are estimated by structured equation modeling, are more similar to traditional models that would be seen with Likert-scale-based measures, with structured factor loadings and uniqueness (Brown & Maydeu-Olivares, 2011), allowing the scores to be better compared between individuals (see Brown & Maydeu-Olivares, 2011, 2013 for an overview of the approach). Indeed, comparisons of traditional and IRT-based scoring of forced-choice measures found the IRT scoring to perform better across multiple types of forced-choice measures in terms of the correlation between the true score and that estimated by the traditional and IRT approaches (Hontangas et al., 2015). However, such approaches are typically focused on fully multidimensional measures due to their ipsative nature, with fewer efforts being focused on scoring approaches for measures that use a combination of unidimensional and multidimensional items.

1.4. Predictive scoring

In contrast to questionnaire-based measures, many contemporary measures of personality use predictive scoring algorithms, using data either from alternative assessments or unstructured data from free text or videos, for example, to predict personality scores on traditional measures (Biel et al., 2012; Hilliard et al., 2022; Kosinski et al., 2013; Leutner et al., 2017; Park et al., 2015). Models of this type, that predict a specified outcome, are said to be supervised (Nasteski, 2017). This is in contrast to unsupervised learning, where algorithms are used to identify clusters in the data, with no specified target variable (see Rosenbusch et al., 2021 for an overview of the types of learning). Since the personality scores are known, these approaches therefore use supervised learning.

While it is possible to use ordinary least squares regression (OLS) to predict scores, predictive measures typically have a large number of predictors, meaning there is a small n/p ratio (Putka et al., 2018). As a result of OLS being designed to minimise the sum of the squared difference between the actual score and predicted score, this can lead to overfitting of the model to the data it was trained on, particularly at small n/p ratios (McNeish, 2015). Since the assessment will be taken by individuals other than those the model was trained on, this can result in the model performing poorly when applied to other samples, limiting its usefulness as a scoring algorithm. Machine learning approaches to prediction can help to overcome this. One approach used in machine learning is least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), which introduces bias into the model, therefore reducing the impact of variance between datasets on the performance of the model due to the bias-variance trade-off (McNeish, 2015). Therefore, compared to OLS regression, Lasso produces a model that is more generalisable to datasets other than the one it was trained on, being more suitable for a scoring algorithm. Another advantage of Lasso regression is that as a result of the regularisation parameter lambda (λ), the coefficient of some predictors is reduced to zero, removing them from the model (McNeish, 2015). Consequently, only the predictors that are most powerful are retained in the model, reducing the complexity of the model and increasing its interpretability (Tibshirani, 1996).

Previously, we reported on the use of Lasso regression to create scoring algorithms for a forced-choice image-based assessment of personality, which presented respondents with pairs of images (items) mapped to the Big Five personality traits and asked them which image in the pair is most like them. After refinement, 150 item pairs (300 images) were retained in the assessment, with these predictors being binarised to represent whether a respondent selected the image or not (Hilliard et al., 2022). All 300 predictors, regardless of which trait they were designed to measure, were entered into the models and used to predict Big Five scores on the IPIP-NEO-120 (Johnson, 2014). This study aims to compare the convergent and discriminant validity, as well as generalisability, of multiple approaches to scoring the image-based assessment:

a) Lasso regression using all 300 predictors,
b) Lasso regression using only the items mapped to each trait (i.e. image pairs that were mapped onto the Big Five traits by the team of I—O psychologists who designed them),
c) OLS regression using all 300 predictors,
d) OLS regression using only the items designed to measure each trait, and.
e) the summative approach to forced-choice measures (again using image pairs as mapped by I—O psychologists).

These models are elaborated on below. Convergent validity will be measured as the correlation between the image-based score and the score on the traditional personality test and discriminant validity will be measured as the inter correlations between the five personality traits and compared to inter correlations on the traditional personality test. Generalisability will be determined by comparing the correlation for the training and test sets, as this can be used to establish how well the model can be applied to unseen data (Jacobucci et al., 2016). It was expected that the machine-learning-based approaches would perform the best in terms of convergent validity and generalisability, followed by the OLS approaches and then the summative approach.

2. Method

Our previously developed image-based measure of personality (Hilliard et al., 2022) presents respondents with pairs of images and asks them to indicate which image in their pair is more like them, with the image pairs, or items, being intended to map on to the statements from the IPIP scales (Goldberg, 1992). Items are either single trait (unidimensional), featuring high and low levels of a single trait, or mixed-trait (multi-dimensional), with the images showing high levels of two different traits (see Hilliard et al., 2022 for an example). The measure, therefore, contains both unidimensional and multidimensional items. Previously, we examined the validity of the measure based on scoring algorithms created using Lasso regression, where all 300 images (150 pairs) were entered as predictors into the model for each trait. This data driven approach was chosen to maximise the predictive validity of the measure. In addition, supervised machine learning approaches can reflect variance from items that contribute to but are not designed to measure a trait since some facets from different traits can be similar (Speer & Delacruz, 2021) (e.g. excitement-seeking from extraversion and adventurousness from openness to experience). This study extends our previous findings, comparing the performance of multiple scoring approaches, including our initial scoring algorithms, in terms of convergent and divergent validity with the IPIP-NEO-120 (Johnson, 2014) and generalisability beyond the training data.

2.1. Participants

For the purpose of this study, we use the same sample as our previous research, namely 431 compensated respondents recruited using Prolific Academic (222 female; 356 under 40 years old; 209 White, 73 Black, 66 Asian, 56 Hispanic, 14 Mixed Race). Respondents took the 150-item image-based assessment along with the IPIP-NEO-120 (Johnson,

2014). The 150 items described in this study were previously selected from a larger item pool based on a sample of 300 compensated respondents (Mage = 31.14, SD = 9.26, 69 % female) who took the IPIP-NEO-120 along with 100 of the image-based items. Based on this sample, the 150 best-performing items were selected based on Cohen's *d* values, which represented the difference in the Big Five scores of respondents selecting image one versus image two (Hilliard et al., 2022), where items with higher Cohen's *d* values performed better and were therefore selected to be retained. These values also enabled the mapping of items to the trait they did measure, instead of what they were intended to measure, by assigning the item to the trait corresponding to the highest Cohen's *d* value (see Hilliard et al., 2022).

2.2. Procedure

In our previous study (Hilliard et al., 2022), we created a separate scoring algorithm for each Big Five trait, with neuroticism being reversed to emotional stability. Specifically, binarised responses to all 300 images were used as the predictor variables and IPIP-NEO-120 scores for the relevant trait as the outcome variable in Lasso models. This approach was selected as the regularisation parameter λ results in some predictors being removed from the model, therefore producing a model with fewer predictors that is more interpretable (McNeish, 2015; Tibshirani, 1996), allowing it to be examined whether the items intended to measure each trait were indeed the most predictive of that trait.

To extend our previous research, in the current study we compare the performance of multiple scoring approaches using both different models (Lasso, OLS and summative) and different combinations of predictors. Since the models are designed to be used as scoring algorithms and therefore need to be generalisable beyond the sample they were trained on, a train-test split was used, with the models being trained on 70 % (n = 323) of the data and tested on the remaining 30 % of the data. This approach allows the generalisability of the models to be examined (Jacobucci et al., 2016) and is common in machine learning since it offers an opportunity for cross-validation. We describe each of the models below.

2.2.1. Lasso models

a) Lasso 300 – regression using all 300 binarised predictors in the model for each trait (Hilliard et al., 2022), with a separate model being created to predict each Big Five trait. This approach was taken as it allowed examination of whether the items designed to measure each trait were predictive of scores for that trait through examining the items retained by each model.

Table 1
Descriptive statistics for the scores for each model (N = 431).

Trait	Mean	SD	Min	Max	Range	Skewness	Kurtosis
IPIP-NEO-120							
Openness	82.84	12.02	34.00	114.00	80.00	−0.32	0.59
Conscientiousness	86.31	15.14	16.00	120.00	104.00	−0.54	1.05
Extraversion	75.98	15.12	11.00	114.00	103.00	−0.30	−0.54
Agreeableness	90.32	13.64	13.00	119.00	106.00	−1.11	3.41
Emotional stability	76.10	18.69	10.00	120.00	110.00	−0.27	−0.15
a) Lasso 300							
Openness	82.89	7.19	59.16	102.31	43.16	−0.05	0.03
Conscientiousness	86.79	10.68	52.17	110.84	58.67	−0.51	−0.12
Extraversion	76.07	11.77	44.2	101.97	57.77	−0.11	−0.65
Agreeableness	90.2	8.46	59.8	110.54	50.74	−0.54	0.51
Emotional stability	75.71	12.67	42.35	100.84	58.49	−0.37	−0.49
b) Lasso intended							
Openness	82.89	4.92	68.93	94.72	25.79	−0.31	−0.29
Conscientiousness	86.57	8.89	63.25	105.09	41.84	−0.40	−0.54
Extraversion	75.95	9.77	53.95	97.52	43.57	−0.05	−0.85
Agreeableness	90.41	6.75	68.22	104.88	36.66	−0.58	0.02
Emotional stability	75.33	10.57	50.84	94.71	43.87	−0.42	0.87
c) OLS 300							
Openness	82.93	11.09	46.75	116.91	70.15	0.01	0.11
Conscientiousness	86.71	14.70	32.97	135.68	102.71	−0.17	0.25
Extraversion	75.84	14.02	33.96	109.69	75.73	−0.27	−0.21
Agreeableness	90.25	11.95	38.71	118.60	79.89	−0.66	0.84
Emotional stability	75.70	16.73	25.67	113.89	88.22	−0.42	−0.17
d) OLS intended							
Openness	82.91	8.14	60.41	102.87	42.46	−0.24	−0.29
Conscientiousness	86.59	11.78	52.27	113.69	61.42	−0.31	−0.26
Extraversion	75.98	12.18	45.25	106.45	61.20	−0.16	−0.46
Agreeableness	90.48	9.91	49.80	113.43	63.63	−0.71	0.67
Emotional stability	75.29	13.12	38.00	103.31	65.31	−0.35	−0.67
e) Summative							
Openness	20.17	6.63	3.00	37.00	34.00	−0.10	−0.32
Conscientiousness	32.35	7.00	9.00	50.00	41.00	−0.30	−0.05
Extraversion	22.37	9.23	3.00	46.00	43.00	0.05	−0.71
Agreeableness	29.58	6.17	9.00	43.00	34.00	−0.33	0.00
Emotional stability	17.12	4.09	5.00	30.00	25.00	−0.13	−0.05

b) Lasso intended - regression using only the items designed to measure each trait (from both the single- and mixed-trait pairs) for each model.

2.2.2. OLS models

c) OLS 300 – OLS regression using all 300 predictors, using the same approach as a), resulting in a low n/p ratio as no predictors were removed from the models.
d) OLS intended – regression using only the items intended to measure each trait (from both the single- and mixed-trait pairs) for each model, resulting in a higher n/p ration than c).

2.2.3. Summative approach

e) Summative – Since the measure contains both single- and mixed-trait pairs (both unidimensional and multidimensional pairs), the traditional summative approach to forced-choice measures was used as this approach would not result in ipsative scores and the popular IRT model (Brown & Maydeu-Olivares, 2011) is only suitable for fully dimensional measures. For the single trait pairs, the positive (high levels) image was assigned a score of 1 and the negative a score of 0. For the mixed-trait pairs, the selected image was assigned a score of 1 and the non-selected 0. Scores were calculated by summing the number of points for each trait.

2.3. Analysis

Model accuracy was determined by correlating the actual and predicted scores (Cui & Gong, 2018) for the training set, while convergent validity with the IPIP-NEO-120 was determined using the correlation for the test set. Although the summative approach is not trained in the same way that a predictive approach is, scores were still grouped into training and test sets to allow for better comparison with the regression models. Further, generalisablity of the model was determined by examining the disparity between the correlation for the training and test sets since this can give insight into how generalisable the models are beyond the training data (Jacobucci et al., 2016).

3. Results

The descriptive statistics for each model can be seen in Table 1, where all predictive models result in a similar distribution of scores regardless of the predictors included in or retained in the model and have a similar mean value compared to the IPIP-NEO-120 scores. Since the number of items designed to measure each trait varies (openness: 49, conscientiousness: 75, extraversion: 67, agreeableness: 62, emotional stability: 47), the maximum score for each trait differs accordingly for the summative scoring approach, resulting in lower mean scores and smaller ranges for the summative models compared to the predictive models. Further, the total score across all five traits for the summative approach ranged from 98 to 140 (M = 121.59, SD = 8.50), demonstrating that the measure is not prone to the issues resulting from ipsative scores that fully multidimensional scales are.

The performance of each model, for both the training and test set, can be seen in Table 2. Since the summative model does not use a regression line, the mean squared error (MSE) and R^2 statistics cannot be calculated using this measure. In general, the Lasso models had the highest convergent validity with scores on the IPIP-NEO-120, with convergent validity for model a) (Lasso 300) ranging from 0.60 for agreeableness to 0.78 for extraversion and model b) (Lasso intended) ranging from 0.58 for agreeableness to 0.77 for extraversion, performing similarly to the prior model. In contrast, the convergence between model c) (OLS 300) ranged from 0.45 for agreeableness to 0.58 for extraversion and model d) (OLS intended) ranged from 0.52 from agreeableness and emotional stability to 0.72 for extraversion. Finally, convergence for model e) (summative) ranged from 0.41 for emotional stability to 0.73 for extraversion. Further, the convergence for the training and test sets for the Lasso models, while lower for the test set, were similar for the training and test sets, indicating that the models are generalisable beyond the data they were trained on. The OLS models, particularly model c), have greater disparity between the correlations for the training and test sets, illustrating how OLS can be prone to overfitting, especially with small n/p ratios. Finally, model e) had no issues with generalisability since it is not a predictive model, although the range of correlations was large and the only model that performed particularly well was the model for extraversion, which has a considerably higher correlation compared to the other traits. Based on these findings, the Lasso models performed the best in terms of both convergent and generalisability, with the full Lasso models (Lasso 300) performing better than the Lasso intended models, followed by OLS models. The OLS intended models performed better than the OLS intended models, likely due to overfitting with the latter model due to the large

Table 2
Performance of each model (N = 431).

	Training (n = 323)			Test (n = 108)		
Trait	r	R^2	MSE	r	R^2	MSE
a) Lasso 300						
Openness	0.77**	0.56	65.63	0.71**	0.50	64.46
Conscientiousness	0.82**	0.66	82.62	0.70**	0.47	97.26
Extraversion	0.86**	0.74	61.23	0.78**	0.61	82.28
Agreeableness	0.77**	0.56	84.97	0.60**	0.34	103.03
Emotional Stability	0.80**	0.63	131.35	0.70**	0.47	175.29
b) Lasso intended						
Openness	0.60**	0.31	102.74	0.68**	0.42	73.93
Conscientiousness	0.71**	0.49	124.05	0.70**	0.48	95.81
Extraversion	0.77**	0.58	98.44	0.77**	0.58	89.52
Agreeableness	0.64**	0.38	121.25	0.58**	0.34	103.10
Emotional Stability	0.66**	0.43	201.86	0.69**	0.44	185.42
c) OLS 300						
Openness	0.83**	0.69	45.91	0.55**	−0.10	141.01
Conscientiousness	0.88**	0.78	53.45	0.46**	−0.46	267.97
Extraversion	0.90**	0.82	42.65	0.58**	0.16	177.97
Agreeableness	0.82**	0.67	64.04	0.45**	−0.18	183.95
Emotional Stability	0.85**	0.73	96.08	0.52**	0.01	327.20
d) OLS intended						
Openness	0.63**	0.40	89.81	0.59**	0.30	89.95
Conscientiousness	0.75**	0.56	106.98	0.66**	0.37	115.10
Extraversion	0.80**	0.64	84.17	0.72**	0.50	104.99
Agreeableness	0.67**	0.45	107.98	0.52**	0.12	137.59
Emotional Stability	0.69**	0.47	187.59	0.66**	0.40	195.51
e) Summative						
Openness	0.34**			0.53**		
Conscientiousness	0.46**			0.52**		
Extraversion	0.66**			0.73**		
Agreeableness	0.52**			0.54**		
Emotional Stability	0.43**			0.41**		

** $p < .001$.

Table 3
Predictors retained by each predictive model and completion time estimates for an assessment with the respective number of (unique) items.

Model	O	C	E	A	ES	Total	Time (mins)
a) Lasso 300	13	26	32	23	30	102	2.5
b) Lasso intended	5	10	19	16	12	62	1
c) OLS 300	300	300	300	300	300	300	5
d) OLS intended	49	75	67	62	47	300	5
e) Summative	49	75	67	62	47	300	5

Table 4
Mulitrait-multimethod correlation matrix for the test set of all models. Convergent correlations with the IPIP-NEO-120 are in bold and the diagonal shows convergent correlations for the training set.

	1	2	3	4	5	6	7	8	9	10	11	12	13
IPIP-NEO-120													
1. O	1.00												
2. C	0.10	1.00											
3. E	0.29**	0.35**	1.00										
4. A	0.31**	0.50**	0.12	1.00									
5. ES	0.01	0.63**	0.54**	0.30**	1.00								
a) Lasso 300													
6. O	**0.71****	0.05	0.39**	0.31**	0.06	(0.77**)							
7. C	−0.04	**0.70****	0.21*	0.32**	0.54**	0.00	(0.80**)						
8. E	0.26**	0.30**	**0.78****	0.12	0.52**	0.48**	0.32**	(0.86**)					
9. A	0.18	0.42**	0.13	**0.60****	0.23*	0.46**	0.54**	0.22*	(0.77**)				
10. ES	0.08	0.51**	0.57**	0.21*	**0.70****	0.19	0.69**	0.71**	0.38**	(0.80**)			
b) Lasso intended													
11. O	**0.68****	−0.17	0.20*	0.03	−0.10	**0.77****	−0.23*	0.25**	0.10	−0.05	.(60**)		
12. C	−0.14	**0.70****	0.20*	0.23*	0.46**	−0.11	0.89**	0.27**	0.38**	0.54**	−0.32**	(0.71**)	
13. E	0.12	0.21*	**0.77****	0.05	0.41**	0.33**	0.18	0.89**	0.12	0.57**	0.17	0.19*	(0.77**)
14. A	0.18	0.33**	0.05	**0.58****	0.18	0.41**	0.47**	0.15	0.93**	0.32**	0.06	0.30**	0.04
15. ES	0.08	0.47**	0.56**	0.15	**0.69****	0.17	0.59**	0.66**	0.20*	0.89**	−0.06	0.50**	0.50**
c) OLS 300													
16. O	**0.55****	−0.07	0.21*	0.25**	−0.05	0.86**	−0.03	0.36**	0.37**	0.13	0.69**	−0.15	0.25**
17. C	−0.05	**0.46****	0.05	0.29**	0.40**	0.02	0.83**	0.20*	0.53**	0.54**	−0.12	0.64**	0.03
18. E	0.21*	0.18	**0.58****	0.12	0.40**	0.48**	0.27**	0.87**	0.27**	0.61**	0.25**	0.15	0.71**
19. A	0.08	0.22*	0.02	**0.45****	0.13	0.35**	0.39**	0.16	0.85**	0.30**	0.12	0.21*	0.08
20. ES	0.16	0.35**	0.43**	0.21*	**0.52****	0.26**	0.51**	0.59**	0.38**	0.86**	0.08	0.30**	0.43**
d) OLS intended													
21. O	**0.59****	−0.12	0.24*	0.02	−0.10	0.72**	−0.24*	0.27**	0.10	−0.05	0.92**	−0.30**	0.23*
22. C	−0.05	**0.66****	0.20*	0.31**	0.44**	0.01	0.88**	0.30**	0.48**	0.56**	−0.23*	0.94**	0.17
23. E	0.12	0.28**	**0.72****	0.12	0.46**	0.35**	0.26**	0.87**	0.21*	0.62**	0.15	0.24*	0.94**
24. A	0.14	0.28**	0.05	**0.52****	0.18	0.38**	0.44**	0.16	0.86**	0.31**	0.07	0.29**	0.03
25. ES	0.03	0.46**	0.50**	0.16	**0.66****	0.14	0.58**	0.61**	0.20*	0.84**	−0.09	0.50**	0.43**
e) Summative													
26. O	**0.53****	−0.37**	−0.19*	−0.11	−0.30**	0.47**	−0.42**	−0.24*	−0.21*	−0.37**	0.71**	−0.50**	−0.34**
27. C	−0.34**	**0.52****	−0.21*	0.00	0.16	−0.45**	0.60**	−0.24*	0.05	0.12	−0.51**	0.71**	−0.30**
28. E	0.06	0.05	**0.73****	−0.10	0.31**	0.23*	−0.03	0.81**	−0.07	0.42**	0.17	0.04	0.92**
29. A	0.12	0.22*	0.06	**0.54****	0.10	0.37**	0.33**	0.18	0.84**	0.25*	0.10	0.22*	0.13
30. ES	−0.15	0.27**	0.15	0.11	**0.41****	−0.23*	0.46**	0.20*	0.05	0.54**	−0.34**	0.37**	0.11

**$p < .001$.
*$p < .05$.

number of predictors. The summative approach performed least well, as only moderate convergent correlations were present for the majority of the traits.

As a result of the regularisation parameter λ, some predictors were removed from the Lasso models. Table 3 shows the number of predictors retained by each model. Since the Lasso models used the fewest predictors but had the highest convergent validity compared to the OLS and summative approaches and were more generalisable, this demonstrates the benefits of using machine-learning-based approaches in predictive measures when measuring personality through alternative formats. When all 300 images were entered as predictors, as can be seen in Table S1, some of the predictors were retained by multiple models, signalling that through using this approach, personality can be rapidly measured through a small number of items that are relevant to multiple traits. In contrast, similar (but lower) convergence is achieved through OLS regression using all 150 items, which would result in a longer completion time. Further, all models contained a mixture of both multidimensional and unidimensional images (see Table S1), retaining the structure of the assessment even when some images were removed from the models and therefore avoiding ipsative scores.

The correlation matrix for the models is shown in Table 4. Using a multitrait-multi-method approach, the convergent and discriminate validity of the measures can be determined (Campbell & Fiske, 1959). There is greater discriminate validity for model b) (Lasso intended) compared to model a) (Lasso 300). Models c) and d) (OLS 300 and OLS intended, respectively) have greater discriminate validity than models a) and b), although this is at the expense of the convergent validity of the models, suggesting that the lower discriminate validity is due to weaker correlations overall. Model e) has the greatest discriminate validity compared to the other models, but also has much lower convergent validity for most traits.

14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
IPIP-NEO-120																
a) Lasso 300																
b) Lasso intended																
(0.64**)																
0.16	(0.66**)															
c) OLS 300																
0.33**	0.07	(0.83**)														
0.46**	0.40**	0.11	(0.88**)													
0.21*	0.50**	0.49**	0.33**	(0.90**)												
0.76**	0.08	0.35**	0.52**	0.31**	(0.82**)											
0.34**	0.68**	0.22*	0.50**	0.61**	0.44**	(0.85**)										
d) OLS intended																
0.04	−0.09	0.68**	−0.15	0.26**	0.10	0.03	(0.63**)									
0.41**	0.50**	0.01	0.72**	0.25**	0.34**	0.37**	−0.24*	(0.75**)								
0.11	0.50**	0.31**	0.15	0.77**	0.17	0.49**	0.20*	0.23*	.(80**)							
0.94**	0.18	0.35**	0.48**	0.26**	0.75**	0.34**	0.03	0.43**	0.10	(0.67**)						
0.16	0.97**	0.04	0.42**	0.49**	0.07	0.67**	−0.12	0.52**	0.45**	0.21*	(0.69**)					
e) Summative																
−0.21*	−0.30**	0.41**	−0.23*	−0.14	−0.12	−0.18	0.56**	−0.40**	−0.31**	−0.16	−0.28**	(0.34**)				
0.00	0.11	−0.39**	0.43**	−0.22*	−0.05	−0.03	−0.46**	0.61**	−0.23*	0.02	0.17	−0.36**	(0.46**)			
−0.10	0.41**	0.14	−0.15	0.58**	−0.08	0.30**	0.24*	0.00	0.79**	−0.09	0.34**	−0.32**	−0.41**	(0.66**)		
0.92**	0.10	0.30**	0.33**	0.22*	0.68**	0.25*	0.08	0.31**	0.17	0.82**	0.09	−0.27**	−0.14	0.01	(0.52**)	
0.03	0.62**	−0.13	0.40**	0.15	0.05	0.38**	−0.37**	0.39**	0.12	0.04	0.57**	−0.38**	0.21*	0.06	−0.06	(0.43**)

Table 5
Model performance for the Lasso and OLS mapped models.

	Training ($n = 323$)			Test ($n = 108$)		
Trait	r	R^2	MSE	r	R^2	MSE
Lasso mapped						
Openness	0.59**	0.31	103.40	0.68**	0.40	76.80
Conscientiousness	0.71**	0.49	123.18	0.68**	0.46	98.53
Extraversion	0.75**	0.55	105.13	0.76**	0.56	93.78
Agreeableness	0.64**	0.37	123.12	0.59**	0.35	101.27
Emotional Stability	0.67**	0.44	198.78	0.67**	0.41	194.88
OLS mapped						
Openness	0.62**	0.39	91.48	0.56**	0.27	92.82
Conscientiousness	0.74**	0.55	110.31	0.64**	0.32	124.57
Extraversion	0.77**	0.59	95.86	0.68**	0.46	114.52
Agreeableness	0.67**	0.45	108.12	0.58**	0.25	117.53
Emotional Stability	0.70**	0.49	181.88	0.63**	0.35	213.49

** $p < .001$.

3.1. Image intention versus image mapping

When refining the measure, we examined whether the item mapped to the trait that it was intended to measure. Based on the Cohen's *d* values calculated in our previous study, 60 % of the items were mapped to the trait they were intended to measure (see Hilliard et al., 2022 for procedure). To examine whether the performance of the models improved when including only the items mapped to the trait, models b) (Lasso intended) and d) (OLS intended) were redeveloped using only the items mapped to each trait using the Cohen's *d* values (Hilliard et al., 2022), instead of the items intended to measure that trait, as determined by the I—O psychologists. As can be seen in Table 5, the convergence of these models is similar to that of models b) and d), with correlations ranging from 0.59 for agreeableness to 0.76 for extraversion for the Lasso mapped model and 0.56 for openness to 0.68 for extraversion for the OLS mapped model. The summative approach was not investigated for the mapped trait as the Cohen's *d* values only indicated that there was a large difference in scores, not whether the image that was designed to be positive was in fact positive for the mapped trait. The coefficients of items for all models can be seen in Table S1.

4. Discussion

This study aimed to compare the performance of machine-learning-based, OLS-regression-based and summative approaches to scoring a forced-choice image-based assessment of the Big Five personality traits. In this section, we compare the performance of each model in terms of generalisability and convergent and discriminate validity with the IPIP-NEO-120. We then provide some potential areas for further research, such as an examination of the predictive validity of the different models for predicting job performance.

4.1. Model evaluation

In this study, we compared the performance of a) Lasso regression using all 300 predictors (150 image pairs), b) Lasso regression using only the images intended to measure each trait, c) OLS regression using all 300 images in each model, d) OLS regression using only the images mapped to each trait in each model, and e) the summative approach to forced-choice assessments. In terms of convergent validity, the Lasso 300 model performed best, followed by Lasso intended, then OLS intended, then OLS 300, and finally the summative model. The two Lasso models performed similarly, and the use of items mapped to each trait, instead of the items intended to measure each trait, did not make a large difference. The OLS intended model performed better than the OLS 300 model, although not as well as the Lasso models. Again, the inclusion of the items mapped to the trait, rather than intended to measure the trait, did not make a big difference on the performance of the model. The worse performance of the OLS models compared to the Lasso models is likely due to the large number of predictors, which would have been a particular issue for the OLS 300 models since they had a much greater number of predictors than the models using only the items intended to measure the trait, likely leading to overfitting (McNeish, 2015). While the summative approach appears to have a large range of convergence values, where the upper end is more in line with the convergence for the other models, this is due to a particularly high correlation for extraversion ($r = 0.73$). In contrast, the correlation for emotional stability is particularly low at $r = 0.41$ and for the remaining traits, convergent validity is around 0.51. Not only is the summative approach less consistent, but the correlations for the remaining traits are also relatively low compared to the correlations for the Lasso and OLS models.

In terms of discriminate validity, the summative approach had smaller correlations than the other approaches and the Lasso models had the largest correlations, therefore following the inverse of the pattern for convergent validity. This suggests that the Lasso models have greater correlations overall, and the summative models have lower correlations. Therefore, none of the approaches we tested can both maximise convergent validity while also maximising discriminate validity. Finally, in terms of generalisability, since the summative models are not predictive, there is no issue with generalisability to unseen data. For the predictive models, generalisability, determined by comparing correlations for the training and test set since the test set acts as an unseen sample (Jacobucci et al., 2016), was greater for the Lasso models than OLS models. However, for the Lasso models, model b) (Lasso intended) had more similar correlations for the training and test sets than model a) (Lasso 300). Models c) (OLS 300) and d) (OLS intended) had lower generalisability, with model c) having a particularly big disparity between the training and test sets, likely due to the small n/p ratio and overfitting of the model.

Due to the generally better performance of the machine-learning-based scoring models, which confirms our hypothesis, we recommend that a machine-learning-based approach to scoring forced-choice measures of personality, particularly those of an image-based format, is a viable option. This finding is in line with previous findings that the convergent validity of forced-choice personality measures is stronger when they are scored using supervised machine learning approaches, as opposed to typical forced-choice scoring approaches (Speer & Delacruz, 2021). As well as the better performance of the machine learning models in our study, Lasso has the additional benefit of removing predictors from the models, leaving only those with the greatest predictive power and resulting in a more interpretable model (Tibshirani, 1996), as well as allowing shorter measures to be derived. This therefore has implications for personality measurement, where more complex assessments based on non-traditional formats or data sources that traditional scoring approaches are not sophisticated enough for can be created and scored using machine learning techniques increasing opportunities for innovation.

4.2. Further studies

Since studies investigating forced-choice, image-based assessments of personality are sparse in the literature, there are a number of directions that future research could take. For example, further studies could more comprehensively examine how the models perform in terms of their discriminant validity relating to other traits since only the IPIP-NEO-120 was used as a point of comparison in this study. Additionally, since other machine learning approaches have been suggested to be appropriate for scoring a forced-choice personality measure, including elastic net regression, deep neural networks, and random forest (Speer & Delacruz, 2021) future studies could compare different machine learning approaches to examine if the performance of the models can be improved further. Alternatively, future research could move away from machine learning based approaches and instead seek to develop an IRT-based scoring approach for unidimensional or mixed dimensional measures since current efforts have focused on IRT approaches for fully multidimensional measures (e.g. Brown & Maydeu-Olivares, 2011, 2013).

Moreover, since the Lasso models reduce the number of predictors retained, (Tibshirani, 1996) further research could examine how these items can be combined to create a measure that rapidly, and accurately, measures personality in around 1 min. These new measures, along with the one described in this study and our previous work (Hilliard et al., 2022), could be examined for predictive validity, using job performance ratings to examine whether the different scoring approaches and combinations of items differ in their ability to predict future job performance. Additionally, future research could address one of the major limitations of the current study, where the measure was only investigated in relation to English-speaking respondents. Although it is claimed that image-based measures do not need to be redeveloped in the target language (Paunonen et al., 1990), the interpretation of image meaning may vary between cultures, meaning that the models may perform differently in other cultures. This could have a particular implication for the Lasso models since the images retained in the models for one language or culture may be less predictive of personality in another culture. Consequently, future research could use a cross-cultural approach to examine the performance of the measure and each model in different cultures.

5. Conclusion

This study supports the use of machine-learning-based scoring models for forced-choice personality assessments, particularly those designed for high-stakes contexts like selection. We found that the machine-learning-based Lasso models performed the best in terms of generalisability and convergent validity, although discriminate validity was weaker than the OLS and summative models. The OLS models had acceptable performance, but resulted in less interpretable models that retained all predictors and were less generalisable to unseen data, likely due to overfitting (McNeish, 2015), particularly with the model using all 300 predictors due to the small n/p ratio. The summative approach performed least well, although it does not have issues with generalisability like predictive approaches can. Based on these findings, we recommend that the best approach to scoring forced-choice personality

measures, particularly if they have an image-based format, is through machine-learning-based predictive scoring algorithms. Based on our findings, we recommend the use of machine learning based predictive scoring algorithms for forced-choice assessments over OLS or summative approaches since machine learning algorithms can maximise the accuracy of the model and generalisability of models to unseen data. Through machine learning, shorter measures can be developed, allowing personality to be measures rapidly through forced-choice statements. While we did not examine whether machine learning can maximise the predictive validity of a measure, our findings show promise for machine learning as a viable scoring method for forced-choice assessments of personality and highlight the possibility for innovative measures of personality to be developed and scored by machine learning.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2022.103659.

Funding

The development of the measure and data collection was funded by HireVue.

Declaration of competing interest

The authors have no conflicting interests to declare.

Acknowledgements

Thank you to everyone involved in the creation of this measure: Sonia-Cristina Codreanu Luca Boschetti, Maurizio Attisani, Clemens Aichholzer, Cari Gardner, and Joshua Liff. This study was conducted using HireVue data and products.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Data availability statement

Data was obtained from HireVue and is not publicly available.

References

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*(1), 1–16. https://doi.org/10.1111/j.1468-2389.2010.00476.x

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Biel, J.-I., & Gatica-Perez, D. (2013). The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia, 15*(1), 41–55. https://doi.org/10.1109/TMM.2012.2225032

Biel, J.-I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012). FaceTube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the ACM International Conference on Multimodal Interaction* (pp. 53–56). https://doi.org/10.1145/2388676.2388689

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52. https://doi.org/10.1037/a0030641

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Costa, P. T., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment: Volume 2 — Personality measurement and testing* (pp. 179–198). SAGE Publications Inc. https://doi.org/10.4135/9781849200479.n9.

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage, 178*, 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001

De Beer, J., Engels, J., Heerkens, Y., & Van Der Klink, J. (2014). Factors influencing work participation of adults with developmental dyslexia: A systematic review. *BMC Public Health, 14*(1), 1–22. https://doi.org/10.1186/1471-2458-14-77

Di Sarno, M., Zimmermann, J., Madeddu, F., Casini, E., & Di Pierro, R. (2020). Shame behind the corner? A daily diary investigation of pathological narcissism. *Journal of Research in Personality, 85*, Article 103924. https://doi.org/10.1016/J.JRP.2020.103924

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. *Computers in Human Behavior, 109*, Article 106356. https://doi.org/10.1016/j.chb.2020.106356

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hilliard, A., Kazim, E., Bitsakis, T., & Leutner, F. (2022). Measuring personality through images: Validating a forced-choice image-based assessment of the big five personality traits. *Journal of Intelligence, 10*(1), 12. https://doi.org/10.3390/jintelligence10010012

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience, 12*, 105. https://doi.org/10.3389/fnhum.2018.00105

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Krainikovsky, S., Melnikov, M. Y., & Samarev, R. (2019). Estimation of psychometric data based on image preferences. In *Conference proceedings for education and humanities* (pp. 75–82). WestEastInstitute. https://www.westeastinstitute.com/wp-content/uploads/2019/06/EDU-Vienna-Conference-Proceedings-2019.pdf#page=75.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences, 49*(4), 331–336. https://doi.org/10.1016/j.paid.2010.03.042

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2021). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000954

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*(1), 113–133. https://doi.org/10.1037/a0021016

Leutner, F., Sonia-Cristina, S.-C., Liff, J., & Mondragon, N. (2020). The potential of game- and video-based assessments for social attributes: Examples from practice. *Journal of Managerial Psychology*. https://doi.org/10.1108/JMP-01-2020-0023

Leutner, F., Yearsley, A., Codreanu, S. C., Borenstein, Y., & Ahmetoglu, G. (2017). From likert scales to images: Validating a novel creativity measure with image based response scales. *Personality and Individual Differences, 106*, 36–40. https://doi.org/10.1016/j.paid.2016.10.007

Lieberoth, A. (2015). Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture, 10*(3), 229–248. https://doi.org/10.1177/1555412014559978

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning, 33*(2), 137–150. https://doi.org/10.1111/jcal.12170

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research, 50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the implicit association test. *Social Psychological and Personality Science, 6*(7), 740–748. https://doi.org/10.1177/1948550615580381

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. In *International conference on*

social computing, behavioral-cultural modeling, and prediction (pp. 48–55). https://doi.org/10.1007/978-3-642-37210-0_6

van de Mortel, T. F. (2008). Faking it: Social desirability response bias in selfreport research. *Australian Journal of Advanced Nursing, 25*(4), 40–48. https://researchportal.scu.edu.au/discovery/delivery/61SCU_INST:ResearchRepository/1267228250002368?l#1367368500002368.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B, 4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Nguyen, L. S., & Gatica-Perez, D. (2016). Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia, 18*(7), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

Park, G., Andrew Schwartz, H., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & 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

Paunonen, S. V., Jackson, D. N., & Keinonen, M. (1990). The structured nonverbal assessment of personality. *Journal of Personality, 58*(3), 481–502. https://doi.org/10.1111/j.1467-6494.1990.tb00239.x

Pletzer, J. L., Oostrom, J. K., & de Vries, R. E. (2021). HEXACO personality and organizational citizenship behavior: A domain-and facet-level meta-analysis. *Human Performance, 34*(2), 126–147. https://doi.org/10.1080/08959285.2021.1891072

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods, 21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass, 15*(2). https://doi.org/10.1111/spc3.12579

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology, 29*(1), 68–74. https://doi.org/10.4102/sajip.v29i1.88

Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. https://doi.org/10.1080/1359432X.2012.716198

Schlosser, A. E. (2020). Self-disclosure versus self-presentation on social media. *Current Opinion in Psychology, 31*, 1–6. https://doi.org/10.1016/J.COPSYC.2019.06.025

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior, 1*(1), 45–65. https://doi.org/10.1146/annurev-orgpsych-031413-091255

Smits, J., & Charlier, N. (2011). Game-based assessment and the effect on test anxiety: A case study. *Proceedings of the European Conference on Games-Based Learning*, 562–566.

Soldz, S., & Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *Journal of Research in Personality, 33*, 208–232. http://www.idealibrary.comon.

Speer, A. B., & Delacruz, A. Y. (2021). Introducing a supervised alternative to forced-choice personality scoring: A test of validity and resistance to faking. *International Journal of Selection and Assessment, 29*(3–4), 448–466. https://doi.org/10.1111/IJSA.12345

Suen, H.-Y., Hung, K.-E., & Lin, C.-L. (2019). TensorFlow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access, 7*, 61018–61023. https://doi.org/10.1109/ACCESS.2019.2902863

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research, 23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Zhang, H., Zhang, J., Sang, J., & Xu, C. (2017). A demo for image-based personality test. *Lecture Notes in Computer Science: MultiMedia Modelling, 10133*, 433–437. https://doi.org/10.1007/978-3-319-51814-5_36