# Image bank creation

The assessment uses a forced-choice format, whereby test-takers are presented with pairs of images and are asked to indicate which image in the pair represents them best. Each pair of images represent different traits from the theoretical model, and are both either desirable traits or undesirable traits. 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) since it is more difficult for test-takers to choose answers that they believe will lead to them being viewed more favourably, which is particularly important in high-stakes contexts like talent management and acquisition.

Items for the assessment were created by subject matter experts to represent the theoretical model. Short statements are also included to support interpretation. The forced-choice pairs represented either two positive sentiments or two negative sentiments such that neither option should be more desirable than the other, giving a truer representation of test-takers instead of being influenced by attempting to choose responses that may be deemed more desirable.

In total, 80 image pairs were created for personality/red flags, and 41 for motivation, totalling 101 questions consisting of 202 images. Accordingly, each personality trait/red flag was presented by 10 images and each motivator by six images.

# Scoring algorithm creation

The assessment is scored using a machine learning based regression approach, where the models are trained to predict personality and motivation scores as measured by validated, questionnaire-based scales using image choices. A machine learning based approach was selected since a summative approach to forced-choice formats, where the number of images designed to measure a certain trait that are selected are summed to get the score for that trait, leads to ipsative scores. This is where everyone receives the same total score across all traits, calculated by adding up the score for each trait, since there is only a finite number of “points” available across all traits (Brown & Maydeu-Olivares, 2013). Methods have been proposed to overcome this based on item response theory (IRT) (Brown & Maydeu-Olivares, 2011, 2013), however, they require specialist implementation and have strict requirements for questionnaire factor structure. Machine learning based approaches, on the other hand, have been demonstrated to perform well with forced-choice image-based assessments of personality (Hilliard et al., 2022b, 2022a) and do not require specialist software or macros to create or run (e.g., Mplus) since they can be directly created in Python. The machine learning based approach also uses a data-driven approach, where the weights of predictors are determined based on patterns in the data, rather than potentially incorrect assumptions about what each item measures. This can vary from the trait that they are designed to measure since images can be more open to interpretation (Hilliard et al., 2022a). Moreover, more datapoints can be used to predict scores since all images (predictors) can be included in the model, which may provide additional insights due to personality traits naturally overlapping.

## Panel data

Models were trained on a dataset that represented global leaders across various industries. The data represents a range of races/ethnicities and nationalities, and includes neurodivergent individuals and individuals with health condition to maximise the representativeness of the data and ensure the algorithm is optimised for different subgroups and ways of thinking.

During the panels, participants were asked to complete the questionnaire-based outcome scales and image-based assessment, as well as self-report their success in their role over the past year.

## Scoring algorithms

The assessment is scored using a machine learning based approach called Ridge regression using image choices to predict scores. Ridge regression has a regularisation parameter that shrinks coefficients in a way that is proportional to their size based on the sum of the squared coefficients, where larger coefficients have greater shrinkage (McNeish, 2015). This can help to make the model more generalisable to other unseen datasets by reducing overfitting. Moreover, Ridge regression is better able to handle collinear predictors, or predictors that correlate highly (McNeish, 2015). This is beneficial for personality assessments where predictors, or image choices, are expected to correlate since respondents will consistently select images that represent their personality traits.

To determine the most appropriate hyperparameters for the models and maximise the generalisability, 10-fold cross validation was used. The models were trained on 70% of the data, with the remaining 30% acting as an unseen sample that could be used to examine the generalisability of the models beyond the training dataset (Jacobucci et al., 2016).

Overall, the models perform well in terms of convergent validity and the performance is generalisable beyond the training data, as determined by the test set correlations. Models also generally demonstrate good discriminant validity, which measures the extent to which an assessment of one construct measures a different construct, where an assessment of one construct should not be strongly related to another construct if the two constructs are theoretically distinct (Hughes, 2017). When creating new assessment measures, a common way to establish the convergent and discriminant validity of the assessment is through the multitrait-multimethod approach, which involves correlating scores of two or more traits across two or more methods of measurement (Campbell & Fiske, 1959). In the case of Napoleon-Bain image-based assessment, this is represented by correlating the scores predicted by the scoring algorithms with scores on the questionnaire-based measures. While there is not a specific threshold for determining whether a measure has discriminant validity, heterotrait-hetero-method correlations should not be as high as convergent validity (Campbell & Fiske, 1959). The personality scores generated by the scoring algorithms also correlated well with self-reported success and have similar predictive validity compared to meta-analytic estimates of the predictive validity of personality traits.

Adverse impact refers to differences in selection rates for different subgroups based on characteristics such as sex/gender and race/ethnicity caused by differences in scores (De Corte et al., 2007). To examine the potential for adverse impact, subgroup scores differnces in scores based on age (binarised into below/above 40 in line with the Age Discrimination in Employment Act), sex/gender, race/ethnicity, presence of a learning difference, and presence of a health condition were examined using three widely used metrics:

* **Four-fifths rule** – compares the pass rates of subgroups to the group with the highest rate to calculate an impact ratio, where ratios below .80 can indicate adverse impact (Equal Employment Opportunity Commission, 1978)
* **Two standard deviations rule** (also known as the z-test) – compares the expected and observed pass rates of each group based on the proportion of data that each subgroup represents, where values >2 indicate that there is a statistically significant discrepancy in expected and observed pass rates (Morgan, 2010; Morris & Lobsenz, 2000)
* **Cohen’s *d*** – a measure of effect size of the difference between means, where values above .20, .50, and .80 indicate small, medium, and large effect sizes, respectively (Cohen, 1992). The current study used a threshold of +/-.30 as indicative of group differences.

Patterns of differences in subgroup scores mirrored the pattern that was observed in the training data, meaning that these group differences were not caused by the assessment itself or the scoring algorithms and were likely to be genuine differences in personality.