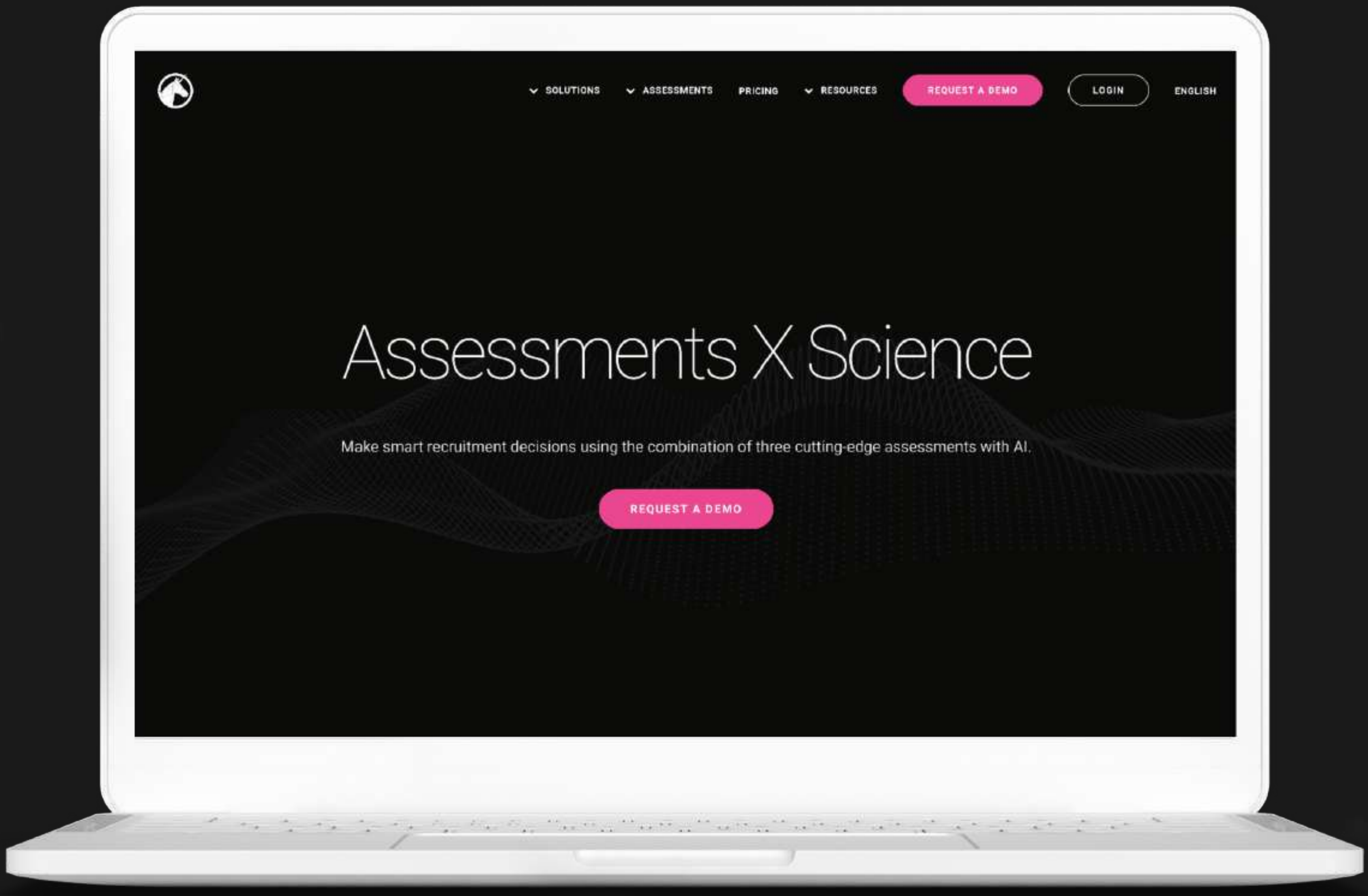


The assessment of potential by AssessFirst

This document provides an overview of the psychometric studies conducted on our personality assessment, SWIPE, our motivations assessment, DRIVE, and our reasoning test, BRAIN. The document outlines the construction of each assessment and presents relevant information regarding their psychometric properties, including validity, reliability, sensitivity, and fairness.

ASSESSFIRST X SCIENCE

Introduction

Making informed HR and Recruitment decisions requires careful consideration and cannot be improvised. Whilst many companies have improved their recruitment practices, some still rely on unreliable methods to screen candidates. For example, the unstructured interview has historically been the most popular tool for this purpose (Buckley, Norris & Wiese, 2000) as it is perceived as more efficient, professional, and natural than other methods (Highhouse, 2008). However, this approach has contributed to deteriorating decision quality by leaving too much room for intuition, prejudices, and cognitive biases (Sinclair & Agerström, 2020; Miles & Sadler-Smith, 2014; Ames, Kammrath, Suppes & Bolger, 2010). As a result, many recruitment processes fail, and discrimination in the selection process persists (Benson, Li & Shue, 2022; Kessler, Low & Sullivan, 2019). Therefore, it is crucial to take the time to measure the attributes that accurately predict a candidate's ability to succeed, instead of giving in to the simplicity and immediacy of intuitive decision-making (Maglio & Reich, 2019; Kirkebøen & Nordbye, 2017). In particular, research in psychology has shown that (1) personality, motivations, and reasoning skills are better predictors of job performance (Sackett, Zhang, Berry & Lievens, 2023; Sackett, Zhang, Berry & Lievens, 2021; Schmidt, Oh & Shaffer, 2016), (2) a simple equation is more efficient and accurate for effective recruitment (Will, Krpan & Lordan, 2022; Kuncel, Klieger, Connelly & Ones, 2013), and (3) companies that follow recommendations from personality and reasoning tests make better hires (Hoffman, Kahn & Li, 2015).

From that perspective, AssessFirst develops and distributes psychometric assessments with the objective of providing HR professionals with reliable indicators of peoples' behavioural attributes. AssessFirst combines tools from behavioural psychology, which have been developed and validated by teams of psychologists and data scientists according to international standards, and AI technology. The compliance of these tools with the standards recommended by the American Psychological Association (APA) and the International Test Commission (ITC) allows AssessFirst to guarantee a high level of quality in the design and continuous improvement of its assessments.

This document provides an overview of the psychometric studies conducted on our personality assessments, SWIPE, our motivations assessment, DRIVE, and our reasoning test, BRAIN. The document outlines the construction of each assessment and presents relevant information regarding their psychometric properties, including validity, reliability, sensitivity, and fairness.

Summary

SWIPE

1. Introduction 6
2. Development history 6
    - 2.1. Fast 6
    - 2.2. Mobile first 7
    - 2.3. Engaging 7
    - 2.4. Reliable 9
3. Theoretical foundations 9
    - 3.1. The Big Five framework and its evolution 9
    - 3.2. SWIPE personality facets 10
4. Development of SWIPE 11
    - 4.1. Phase 1: testing series 11
    - 4.2. Phase 2: item selection 13
    - 4.3. Phase 3: validation series 13
5. Final version 14
6. Validity 15
    - 6.1. Content validity 15
    - 6.2. Construct validity 21
    - 6.3. Convergent validity 29
    - 6.4. Predictive validity 30
    - 6.5. Conclusion 31
7. Reliability 32
    - 7.1. Internal consistency 32
    - 7.2. Test-retest reliability 41
8. Sensitivity 42
9. Fairness 43
    - 9.1. SWIPE accessibility 43
    - 9.2. Fairness in SWIPE results 44

DRIVE

1. Introduction 50
2. Development history 50
3. Theoretical foundations 51
    - 3.1. Self-determination theory 51
    - 3.2. Motives, Values, Preferences Inventory 52
    - 3.3. Multidimensional theory of person-environment fit 52
    - 3.4. DRIVE's dimensions 53

4. Development of DRIVE 54
- 4.1. Choice of the dimensions 54
- 4.2. Items' conception 55
- 4.3. Development 55
- 4.4. Format 57

5. Validity 58
- 5.1. Content validity 58
- 5.2. Construct validity 60
- 5.3. Predictive validity 63
- 5.4. Conclusion 64

6. Reliability 64
- 6.1. Internal consistency 64
- 6.2. Test-retest reliability 65

7. Sensitivity 66

8. Fairness 67
- 8.1. Fairness in DRIVE's results 67

BRAIN

1. Introduction 71

2. Development history 71

3. Theoretical foundations 73
- 3.1. Theoretical elements 73
- 3.2. Type of analysis available in BRAIN 74

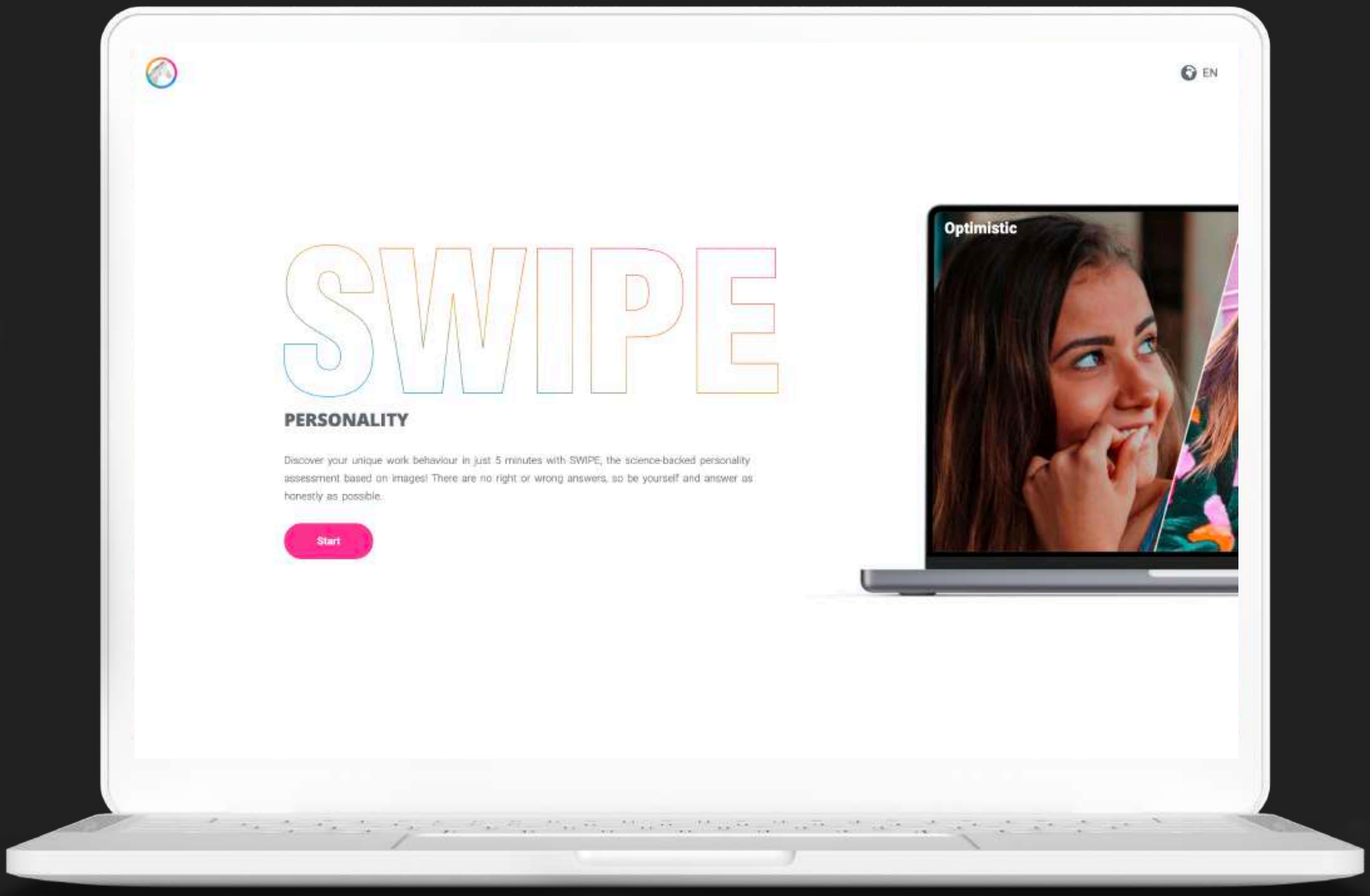
4. Development of BRAIN 75
- 4.1. Phase 0 75
- 4.2. Phase 1 75
- 4.3. Phase 2 75
- 4.4. Phase 3 76
- 4.5. Phase 4 76

5. Validity 77
- 5.1. Construct validity 77
- 5.2. Convergent validity 79
- 5.3. Predictive validity 81
- 5.4. Conclusion 81

6. Reliability 82

7. Sensitivity 82

8. Fairness 85
- 8.1. Fairness in BRAIN's results 85



Welcome to the future of personality assessment

SWIPE is a short, image-based personality assessment that provides insight into how an individual behaves in a professional setting. With a mobile-first design and a duration of only 5 minutes, SWIPE incorporates the latest in psychometric research and user experience.

SWIPE

1. Introduction

SWIPE is a short, image-based personality assessment that provides insight into how an individual behaves in a professional setting. With a mobile-first design and a duration of only 5 minutes, SWIPE incorporates the latest in psychometric research and user experience. The assessment consists of 72 items that measure 6 traits and 18 personality facets, along with 3 data collection items, for a total of 75 items. The AssessFirst Science team developed SWIPE in 2023, and it has already been the subject of 6 papers presented at international psychology conferences or published in peer-reviewed scientific journals.

2. Development history

Whilst personality is a determining factor in predicting success in the workplace (Judge & Zapata, 2015), the personality assessments currently available on the market are often considered lengthy, outdated, and do not provide a good user experience. As a result, traditional assessments usually receive average favourability scores (Hausknecht, Day & Thomas, 2004). These conclusions seem reasonable in a world where everything is becoming faster, more visual, and mobile-first: Instagram for sharing photos, Spotify for listening to music, Google Maps for finding your way with a few clicks, or Tinder for finding love whilst swiping are all examples of the new ways of accessing information and how we interact with digital technology and its capabilities. Furthermore, in response to the increasing desire of candidates for faster and more accessible recruitment processes via smartphones (Böhm & Jäger, 2016), companies must adopt these new standards to remain competitive and attractive. It is in light of these observations and the ever-growing needs of HR professionals to design decision-making processes that are fast, reliable, and engaging, that SWIPE was developed (Kubiak, Niesner & Baron, 2023). Specifically, the development of SWIPE was driven by four needs or assumptions.

2.1. Fast

The ideal length of a personality assessment is a complex topic that requires consideration of candidate expectations, the perception of fairness and seriousness of the measurement, and the validity of the assessment. The objective is not necessarily to propose an assessment that is as quick as possible, but rather to find the perfect balance to optimise our response to these needs. Scientific studies have concluded that (1) candidates prefer an overall assessment time between 10 and 30 minutes and that the majority of applicants who quit assessments did so within the first 20 min of the assessment phase (Hardy, Gibson, Sloan & Carr, 2017), (2) assessments that are too long can lose validity (Burisch, 1997) and the measure can be influenced by other factors (Myszkowski, Storme, Kubiak & Baron, 2022), (3) assessments that are too short, although useful, do not capture all the information about a person's personality (Hofmans, Kuppens & Allik, 2008), (4) the ideal number of data points per scale is between 6 and 9 measurements (Soto & John, 2019). For the development of SWIPE, we aimed to maximise these elements by building an assessment with 8 main items per facet and lasting an average of 5 minutes, which would result in an overall assessment time of approximately 25 minutes, taking into account the DRIVE and BRAIN assessments.

2.2. Mobile first

The way candidates prefer to complete assessments as part of a recruitment process has significantly changed in recent years and is moving towards mobile usage (Lawrence & Kinney, 2017; Smith, 2015). This method of administration has several advantages, such as: (1) responding to societal and technological developments, where mobile devices are the main tool for media consumption (Goovaerts, 2016) and internet access (Smith, 2015), (2) allowing candidates to complete assessments anywhere and anytime (Arthur & Traylor, 2019), and (3) increasing accessibility to historically discriminated groups such as women, African-American, and Hispanic populations who are more likely to complete mobile-based assessments (Arthur, Doverspike, Muñoz, Taylor, & Carr, 2014). However, it is naive to simply convert a computer-based personality assessment to a mobile device without ensuring its full adaptation and accessibility as it may negatively impact the user experience (Gutierrez & Meyer, 2013). Instead, assessments designed for deployment on mobile devices, from the outset, offer a consistent user experience, regardless of the device being used (Kinney, Lawrence, & Chang, 2014). Therefore, SWIPE has been designed mobile-first with our team of psychologists and UX designers to optimise the user experience.

2.3. Engaging

Whilst SWIPE's speed and mobile-first design help to make it more engaging and appreciated by users, three main elements are at the heart of SWIPE's quality of experience, namely:

- **Gamification through images**: Whilst personality is often measured using assessments with classic Likert scales, new trends and technological possibilities are making gamification a lever for user engagement (Leutner, Akhtar, & Chamorro-Permuzic, 2022; Leutner & Chamorro-Permuzic, 2018; Armstrong, Ferrell, Collmus, & Landers, 2016; Chamorro-Permuzic, Winsborough, Sherman, & Hogan, 2016). In this sense, gamified evaluations are perceived as more immersive than traditional evaluations (Leutner, Codreanu, Liff, & Mondragon, 2020), reduce user anxiety (Mavridis & Tsiatsos, 2016), and increase user satisfaction, resulting in a stronger perception of the fairness of the recruitment process and better organisational attractiveness when these assessments are used (Georgiou & Nikolaou, 2020). Among the different means of gamification, the use of images to measure personality has proven to be an effective strategy (Hilliard, Kazim, Bitsakis, & Leutner, 2022; Leutner, Codreanu, Liff, & Mondragon, 2020; Krainikovsky, Melnikov, & Samarev, 2019; Leutner, Yearsley, Codreanu, Borenstein, & Ahmetoglu, 2017), allowing for both valid measurement of personality (Kubiak, Niesner, & Baron, 2023) and optimisation of user satisfaction (Efremova, Kubiak, & Baron, 2023). Indeed, using images provides more context and information to the user, making it easier and faster to read and process compared to text (Potter, Wyble, Hagmann & McCourt, 2014). Additionally, images can provide additional data points that can help infer the personality of the respondent (Kubiak, Bernard & Baron, 2023). However, SWIPE goes beyond just using images by drawing inspiration from research in marketing, consumption, and decision sciences. Studies have shown that hybrid formats, combining a short text with an image, are more effective (Wu, Wu & Wang, 2020), that images can overcome biases linked to the fact that people don't read long texts (Zinko, Stolk, Furner & Almond, 2019), and that professional-quality images that represent humans and have optimal text-image associations can increase user engagement on social media (Lie & Xie, 2019). Capitalising on these findings, SWIPE is designed as an "image-based" assessment, where each image is accompanied by a short descriptive text to provide high-quality information and high user engagement.

- **The « swipe » as a means of response**: The emergence of mobile consumption has also contributed to integrating new means of physical interaction with information. Among these, the "swipe" - a touch of the screen followed by a sliding movement, has established itself as one of the most used gestures by mobile application designers and has become a part of our daily lives. The "swipe" simplifies actions and decisions on mobile by making them binary, allowing things to be done more quickly (Rodrigues & Baldi, 2017). It also proves to be more fluid, intuitive, and understandable for users, thereby increasing their satisfaction (Dou & Sundar, 2016). In short, the logic of the swipe, if it is primarily technical and physical, also serves as a lever of satisfaction and psychological persuasion (David & Cambre, 2016). In the field of personality assessment, new research has highlighted the beneficial effects of swiping for user engagement (Efremova, Kubiak & Baron, 2023) and reduced response time per item (Weidner & Landers, 2020). Whilst users can answer the SWIPE assessment using different means, especially in the desktop version, the swipe movement is the only way to complete the assessment in its mobile version.

- **Forced-choice format**: The forced-choice response format, where the respondent must choose between two options, is gaining popularity and positioning itself as an alternative to Likert-type or single-statement measures. Forced choice helps neutralise acquiescence, extremity, or self-serving biases (Wetzel, Böhnke & Brown, 2016; Paulhus & Vazire, 2007), and can drastically reduce cheating attempts (Cao & Drasgow, 2019). Coupled with IRT scoring models (Brown & Maydeu-Olivares, 2011), forced-choice response formats are therefore more effective in measuring personality. However, this response format can be cognitively heavier for users, leading to more complicated decision-making and a less satisfactory experience (Bartram & Brown, 2004). To benefit from the advantages of forced choice whilst improving the user experience, adaptive actions are required to overcome the mixed reactions inherent in this format. SWIPE takes into account three types of fixes that have demonstrated their effects in improving user engagement with this response format (Dalal, Zhu, Rangel, Boyce & Lobene, 2021):

  - (1) One criticism of forced-choice formats is that users may wish to select both options or neither, and the absence of this possibility in current assessments can lead to frustration (Bartram & Brown, 2004). To address this issue, SWIPE will give users the option, a certain number of times, to select both answer options or neither. Our studies have shown that this double-choice option improves the user experience (Efremova, Kubiak & Baron, 2023) and provides valuable information on the respondent's personality (Baron, Storme, Myszkowski & Kubiak, 2023; Myszkowski, Storme, Kubiak & Baron, in press);

  - (2) Include feedback after completing the assessment: A summary is generated automatically and presented to the respondent after completing SWIPE. This summary provides the respondent with concrete elements of understanding their personality, preferred behaviours, personal style, and areas for improvement. All the content is positively phrased and aims to help respondents get to know themselves better in a simple and objective way. According to our studies, 90% of users find this summary easy to understand, and 98% find it useful[2] ;

  - (3) To ensure the good validity of the assessment, a balance between positively and negatively formulated items is necessary (Soto & John, 2019). However, we also took care to remove items and response proposals that were considered too negative and thus never selected by the respondents. This strategy avoids including items that may require the respondent to make a complex or psychologically embarrassing choice.

[2] Qualitative survey conducted among 180 users.

2.4. Reliable

Among all the existing personality frameworks, the Big Five model has, for many years, through several thousand studies, demonstrated its validity, reliability, and usefulness (Goldberg, 1993b; John, Naumann, & Soto, 2008; McCrae & Costa, 2008). Notably, the Big Five personality facets have consistently predicted job performance (Barrick & Mount, 1991; Judge, Higgins, Thoresen & Barrick, 1999; Higgins, Peterson, Pihl & Lee, 2007; Kuncel, Ones & Sackett, 2010; Schmitt, 2014), especially when contextualised to the requirements of a specific profession (Judge & Zapata, 2015; Tett, Toich & Ozkum, 2021). The popularity of this model leads the scientific community to constantly challenge and improve it: research on the "Big Five Inventory-2" has introduced a more robust hierarchical structure to the model, improved its fidelity and predictive power, and retained the original model's conceptual orientation and ease of understanding (Soto & John, 2017a). Recently, the BFI-2 model has further evolved: this model has historically been considered sub-optimal for evaluating the Honesty-Humility (H) scale of the HEXACO. Still, new research has proposed the addition of three ad hoc facets for measuring this H scale, thus improving the BFI-2's measurement of the Honesty-Humility dimension (Denissen, Soto, Geenen, John & Van Aken, 2022; Lee, Ashton & De Vries, 2022). To ensure that SWIPE is based on the most effective and modern personality models, it has been designed to maximise convergent validity with the BFI-2 and its new Humility scale. More details are provided in the next section.

3. Theoretical foundations

3.1. The Big Five framework and its evolution

The SWIPE personality assessment is based on the Big Five model (Goldberg, 1993b; John, Naumann, & Soto, 2008; McCrae & Costa, 2008). This model, also known as the Five Factor Model (FFM), was initially developed through a factor analysis of a large number of evaluation reports on adjectives and personality assessment items. From a lexical perspective, the development of FFM (Digman, 1990; Goldberg, 1992; John, 1990; McCrae & Costa, 1987) is based on several decades of research. The Big Five model identifies five major personality traits: Extraversion, Agreeableness, Openness to Experience (or Openness), Conscientiousness, and Emotional Stability (also called Neuroticism). Details on each of these traits can be found in Table 3.1.

Traits	Description	ACL marker items
Extraversion	The degree to which the person seeks social interactions with others.	Quiet, Reserved, Shy vs. Talkative, Assertive, Active
Agreeableness	The degree to which the person cultivates harmonious relationships with others.	Fault-finding, Cold, Unfriendly vs. Sympathetic, Kind, Friendly
Openness	The degree to which the person pursues intellectual challenges and exhibits curiosity.	Commonplace, Narrow-interest, Simple vs. Wide-interest, Imaginative, Intelligent
Conscientiousness	The degree to which the person conforms to social norms and standards.	Careless, Disorderly, Frivolous vs. Organised, Thorough, Precise
Emotional stability	The degree to which the person experiences negative emotions.	Tense, Anxious, Nervous vs. Stable, Calm, Contented

Table 3.1. Big Five traits and their description.

This model provides an excellent foundation for the development of personality assessment tools. In fact, studies have shown that all multidimensional personality inventories can be reorganised around these five major traits (Raad & Perugini, 2002). In other words, all inter-individual differences in behaviour, feelings, and ways of thinking can be summarised by these five traits. Since its inception, this model has gained strong scientific recognition and robustness. Several decades of research have contributed to its refinement and to the development of validated measurement assessments, such as the "Big Five Inventory" (BFI), which consists of 44 Likert-type items (John, Donahue, & Kentle, 1991).

However, in the 30 years since the creation of the BFI, our understanding of personality, its structure, and its evaluation has been refined and improved. Recent research has integrated this new knowledge whilst addressing the structural and psychometric limitations identified in the first version of the BFI. This has resulted in the publication of the "Big Five Inventory - 2" or "BFI-2" (Soto & John, 2017a), which introduces a more robust hierarchical structure to the model consisting of 15 personality facets. The BFI-2 improves the fidelity and predictive power of the model whilst retaining the original conceptual orientation and ease of understanding. The BFI-2 is composed of 60 Likert-type items, but shorter versions have also been developed, including the "BFI-2-S" consisting of 30 items and the "BFI-2-XS" consisting of 15 items (Soto & John, 2017b). The inventory has also recently been adapted into French and was used in the development of SWIPE (Lignier et al., 2022). These results demonstrate that the BFI-2 is a reliable and valid measure of the Big Five traits and their associated facets and that it represents a significant advance over the original BFI version.

However, some studies suggest that the "BFI" and "BFI-2" scales may be limited in capturing the variance related to humility, which is an important personality trait, particularly in a professional context. The correlations between the Big Five and the "Honesty-Humility" sphere of the HEXACO-PI-R seem to be weaker for the "BFI-2" compared to other inventories, such as the NEO-PI-R (Ashton, Lee, & Visser, 2019; Ashton & Lee, 2019; Lee & Ashton, 2019). Recent analyses support these conclusions, highlighting the "BFI-2's" limited ability to account for the "H" scale associated with humility. Given the impact of this scale on predicting pro-social behaviors (Thielmann, Spadaro, & Balliet, 2020), it is necessary to extend the Big Five and the "BFI-2" model by integrating a related trait for humility. New research proposes adding three ad hoc facets to the "BFI-2" model to measure the H scale (Denissen, Soto, Geenen, John, & Van Aken, 2022; Lee, Ashton, & De Vries, 2022).

3.2. SWIPE personality facets

The Big Five have consistently demonstrated their ability to predict success in the workplace and various aspects of everyday life (Soto, 2019; Soto, 2021). With this in mind, SWIPE was developed to maximise convergent validity with the "BFI-2" and its newly added humility scale. To provide a more detailed analysis of each profile, a facet approach is preferred. In simple terms, a personality facet is a distinct pattern of thought, feeling, or behaviour that tends to remain stable across different situations and over time (Allport, 1961; Bleidorn et al., 2022). As a result, SWIPE measures 6 traits and 18 personality facets, which are presented and defined in Table 3.2.

Traits	Facets	Description
EXTRAVERSION	Assertiveness	Tendency to behave as a leader and influence others
	Energy level	Tendency to show enthusiasm and energy
	Sociability	Tendency to approach others easily, be sociable, and extroverted
AGREEABLENESS	Compassion	Tendency to be benevolent and compassionate towards others
	Respectfulness	Tendency to be respectful, polite, and avoid conflict
	Trust	Tendency to easily trust and forgive others
HUMILITY	Greed avoidance	Tendency to focus on simple things, be unmaterialistic
	Modesty	Tendency to show modesty and humility
	Sincerity	Tendency to show sincerity and be honest
OPENNESS	Aesthetic sensitivity	Tendency to be interested in art in all its forms
	Creative imagination	Tendency to be inventive, creative, and original
	Intellectual curiosity	Tendency to be curious and interested in abstract things
CONSCIENTIOUSNESS	Organisation	Tendency to organise methodically and be methodical
	Productiveness	Tendency to seek maximum performance and be efficient
	Responsibility	Tendency to be reliable and respect commitments
EMOTIONAL STABILITY	Anxiety	Tendency to feel stress and be reactive
	Depression	Tendency to experience predominantly negative emotions
	Emotional volatility	Tendency to express and share one's emotions and feelings

Table 3.2. Personality facets measured by SWIPE.

4. Development of SWIPE

The development of SWIPE took place via a 3-phase process: (1) the creation and testing of several sets of items, (2) the selection of the best-tested items, and (3) the creation and testing of a validation series composed of the selected items. These three steps are presented below in more detail.

4.1. Phase 1: testing series

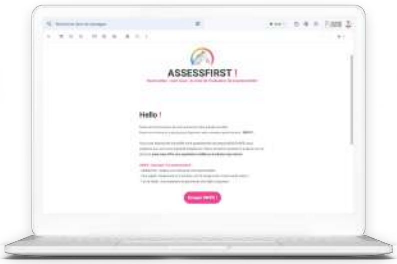
Between May and December 2022, six sets of test items were created and released with the aim of collecting data on as many SWIPE items as possible, which were supposed to measure personality facets of the BFI-2. This was important because not all items tested proved to effectively measure the facet they were intended to measure. The series was launched sequentially, and when enough data had been collected on series 1, series 2 was put into production, and so on. In order to collect high-quality data and maximise respondent response rates, each series was composed of only 60 items in total, with each item consisting of a pair of images, each associated with a short textual description. A total of 360 items (i.e., 720 single-answer choices) were tested to develop SWIPE. These items were constructed by a team of 5 psychologists, drawing from the theoretical and semantic corpus linked to the Big Five.

The respondents of these series were invited to participate in the study if they met three essential conditions: (1) they had created an AssessFirst account in 2022, (2) they had completed the SHAPE assessment, and (3) they had consented to receive commercial and scientific communication from AssessFirst in compliance with the GDPR. To maximise the response rate, several email reminders were sent.

Finally, to study the item quality and convergent validity of SWIPE with respect to the BFI-2 and its humility scale, participants who completed a series of SWIPEs were invited to also complete the version of the BFI-2 (Lignier et al., 2022) and an additional 12 items measuring the three facets of the humility domain (Denissen et al., 2022). This second assessment consisted of 72 Likert-type items and was accessible on Typeform after having completed the SWIPE series.

In summary, participants who agreed to participate followed the procedure outlined below:

Step 1
INVITATION



Receipt of the invitation email presenting the data collection process and its purposes.

Step 2
SWIPE

Completion of a series of SWIPE, consisting of 60 items that took about 5 minutes to complete.

Step 3
BFI-2 + H Scale

Completion of the BFI-2 and humility scale via the Typeform platform, that took approximately 10 min.

In total, the first phase of the study involved 2,989 participants, with approximately 500 respondents per series. The descriptions of the respondent samples are presented in Tables 4.1 and 4.2 below:

Series	Respondents	Gender			Work experience
	Total	Female	Male	Non-binary	Mean (in years)
01	501	61 %	37 %	2 %	9,8
02	483	63 %	36 %	1 %	10,9
03	541	56 %	43 %	1 %	10,2
04	458	62 %	37 %	1 %	10,4
05	497	65 %	34 %	1 %	11,3
06	509	58 %	41 %	1 %	10,8
Total	2989	61 %	38 %	1 %	10,6

Table 4.1. Description of the gender and average work experience of the respondents for the series phase.

Series	Diploma					
	PhD	Master	Bachelor	A-Level	Occupational	None
01	2 %	36 %	27 %	15 %	14 %	6 %
02	3 %	42 %	31 %	13 %	8 %	3 %
03	3 %	38 %	30 %	12 %	12 %	5 %
04	1 %	37 %	33 %	12 %	12 %	5 %
05	3 %	37 %	29 %	15 %	13 %	3 %
06	2 %	33 %	33 %	14 %	13 %	5 %
Total	2 %	37 %	30 %	14 %	12 %	5 %

Table 4.2. Description of the educational level of the respondents for the series phase.

These statistics support the diversity of users who participated in the first phase of the development of SWIPE. Although the results in terms of diploma or education level distribution show a slight over-representation of users with a Bachelor's or Master's degree, this can be explained by two main reasons: (1) the natural and increasing representativeness of people with tertiary education levels in the population - around 50% according to OECD figures from 2017 to 2020, and (2) the slightly increased use of the AssessFirst solution for recruiting executives profiles by our clients - explaining their prevalence in our contact database.

4.2. Phase 2: item selection

The data collected in each series were analysed by our team of psychologists to choose the items and response choices that best measured their reference construct. These items were selected based on several rules and conditions, including:

- A good correlation with the reference construct ($r > .30$ or $r < -.30$);
- Both answer choices of an item fulfilling the previous condition;
- Optimal content validity with the construct;
- Avoiding semantic repetition for the same facet;
- Balancing "positive" and "negative" answer choices for the same facet;
- Diversifying items, with the ideal goal of having a unique item for each combination of 2 facets;
- Including a diverse range of main characters in the images, in terms of gender and origin.

Following the selection process, we were able to identify 135 unique items out of the 360 tested across the 6 proposed series. These 135 items underwent further data collection in a new phase.

4.3. Phase 3: validation series

In order to select the final items for SWIPE, the 135 chosen items were subjected to a validation series to collect information on each item within a larger sample. Respondents who met three essential conditions were invited to participate in this final series: (1) having created an AssessFirst account in 2022, (2) having completed the SHAPE assessment, and (3) having accepted commercial and scientific prospecting communications by AssessFirst in compliance with GDPR. To maximise the response rate, several email reminders were sent, and some Phase 1 participants were also re-invited to complete the final series. This data collection occurred from 09.01.23 to 10.02.23. As in the first phase, participants were asked to complete SWIPE and then the BFI-2 inventory.

Series	Respondent	Gender			Work experience
	Total	Female	Male	Non-binary	Mean (in years)
Validation	4457	54 %	33 %	13 %	10.9

Table 4.3. Description of the gender and average work experience of the respondents for the validation phase.

Series	Diploma					
	PhD	Master	Bachelor	A-Level	Occupational	None
Validation	2 %	39 %	31 %	13 %	11 %	4 %

Table 4.4. Description of the educational level of the respondents for the validation phase.

Once again, the collected data was analysed by our team of psychologists to select the items that best measured their reference constructs. The items were first sorted based on the same criteria as in Phase 2. Then, the selection of items for the industrial version of SWIPE was done by testing several statistical models. The main characteristics of the selected items are presented in the following section.

5. Final version

75 The final version of SWIPE is composed of 75 forced-choice items. Out of these 75 items, 3 are used for data collection purposes.

6 Number of personality traits covered by SWIPE, allowing for a comprehensive assessment of personality.

18 Number of personality facets assessed by SWIPE. These facets are derived from the BFI-2 and an additional humility scale.

5 The average response time. On average, respondents take 4.19 seconds to answer an item.

6. Validity

How can we determine if an assessment accurately measures what it claims to measure? How can we ensure that each scale is measured correctly and that the results of the assessment have the intended meaning? These questions are answered through validation studies. The purpose of validating an assessment is to confirm that it measures the intended construct and to determine the accuracy of the results obtained from it. In the past, validity was defined as the correlation between a score on an assessment and an external criterion that measured either the same construct or a construct that was supposed to be related to the construct associated with the score. To establish and ensure the validity of an assessment, several types of validity must be examined. The validity studies of SWIPE cover the following types of validity:

- **Content validity**: refers to the extent to which the items of an assessment semantically represent an adequate sample of the content domain being measured. This means that the items should be directly related to the construct they are intended to measure and also cover all the main aspects of that construct;
- **Construct validity**: refers to the degree to which the assessment accurately measures the psychological construct or facet it is designed to assess. This type of validity is established through various analyses, such as item-dimension saturation, inter-dimension correlation, RMSEA, and distribution parameters;
- **Convergent validity**: refers to the degree to which two measures of constructs that should theoretically be related are indeed related. In other words, convergent validity measures the degree to which the results of one assessment are correlated with those of another assessment that assesses the same or a similar concept;
- **Predictive validity**: the predictive validity of a personality assessment measures its ability to predict a target variable, such as job performance or turnover. In other words, it assesses whether the results of the personality assessment can be used to predict future outcomes in the workplace.

6.1. Content validity

6.1.1. Introduction

Content validity assesses the relevance of the content of an assessment by examining whether it represents all facets of a given construct and whether its items are representative of the construct being measured. This type of validity is crucial because the development of items for a personality assessment is primarily a trial-and-error process (Tellegen & Waller, 2008). If the items are poorly developed, the scales measured by the assessment may not adequately represent the construct being measured (Smith, Min, Ng, Haynes & Clark, 2022). Content validity allows researchers to assess the extent to which an item's content is related to the personality construct it is intended to measure (Worthington & Whittaker, 2006; Colquitt, Sabey, Rodell & Hill, 2019). Historically, content validity has relied on a rational approach to linking item content to the construct rather than statistical analysis. A common approach to assess content validity is to solicit expert judges who will evaluate the relevance of items through a manual exercise of item classification. Several indicators are then calculated, such as inter-judge agreement, which represents the proportion of judges who indicate that the item is semantically well linked to the construct it measures (Anderson & Gerbing, 1991; Fleiss, 1981). However, this approach is subject to several limitations as it is time-consuming and cognitively costly (Krippendorff, 2018; Short, McKenny & Reid, 2018). Additionally, the expertise of the judges selected may influence the results, and they may be imprecise in their classifications (Fyffe, Lee & Kaplan, 2023).

Machine learning (ML) and natural language processing (NLP) techniques are increasingly being applied in behavioural science for content creation and analysis (Campion, Campion, Campion & Reider, 2016; Hommel, Wollang, Kotova, Zacher & Schmukle, 2022; Jiao & Lissitz, 2020; Lee, Fyffe, Son, Jia & Yao, 2023; Von Davier, 2018). These technologies can overcome the limitations of traditional content validity methods and significantly optimise the validity process. Recent research by Fyffe, Lee, and Kaplan (2023) proposes a new approach based on NLP and transformer models. Transformers are a type of deep neural network that converts text into digital representations. Unlike previous natural language models that mainly used recurrent neural networks (RNNs), transformers rely on a parallel processing architecture that better considers the relationships between the different elements of the text sequence. By applying transformers to text classification, researchers have developed an automated approach to content validation of personality scales. Compared to the traditional approach previously described, this method reduces procedural and cognitive complexity whilst optimising classification performance. This methodology represents a major advancement in the ability of publishers to build effective measurement scales and validate content quality.

6.1.2. How does it work?

Classification tasks involve training a model to categorise text into predefined categories. A classification model is, therefore, a type of machine learning model that is used to predict the class or category of an object or observation based on its characteristics. In the context of the SWIPE assessment, the first step is to train a classification model that can determine the personality trait associated with each item. The development of this classification model involves four main steps:

- Creating a training dataset: To build an effective classification model, it is necessary to collect a dataset that contains personality items and the Big Five trait to which they belong. These data must be representative of the different classes that we want to predict. Additionally, the quality of the data should be ensured as the classification algorithm learns on the basis of this data;
- Textual representation: This step involves encoding textual data (items) as digital vectors that can be processed by machine learning algorithms. As machine learning algorithms cannot directly process plain text, it is necessary to encode them as digital vectors. This numerical representation takes into account important characteristics of the text, such as the words used, their order, their frequency, etc;
- Model training: consists of teaching a classification algorithm to identify the relationships between the input characteristics (the items) and the output variable (the personality trait to be predicted). In other words, the goal of training the model is to find a function that relates the input features to the output class. This is done by providing the algorithm with a labeled dataset and iteratively adjusting the model's parameters until it can accurately predict the correct class for new, unseen items;
- Model evaluation: After training the model, it is evaluated using a neutral sample. Different performance metrics are used to assess its quality, such as accuracy, precision, recall, and F1-score.

6.1.3. The AssessFirst classification model

In order to build upon the existing wealth of scientific and open-source literature on this subject, our studies draw from certain results already obtained by Fyffe, Lee, and Kaplan (2023). Based on their findings, we have made several choices that allow us to better meet our objectives.

On the one hand, whilst their studies were able to train a classification model to predict the five traits of the Big Five, our SWIPE studies needed to go further. In addition to the Big Five, SWIPE includes a dimension of humility. To address this, our team selected a new training dataset that included items related to the humility trait, in addition to the dataset used by Fyffe, Lee, and Kaplan (2023). The items were selected from reputable assessments (such as BFI, BFI-2, BFI-10, HEXACO-100, HEXACO-60, HEXACO-24, BFAS, and NEO-PI-R) and open-source databases (such as the International Personality Item Pool or IPIP). By enriching the dataset with these items, we were able to train the model to predict the class of humility, without compromising the quality of the data or the performance of the model.

Trait	Number of items
Extraversion	669
Agreeableness	762
Humility	246
Openness	777
Conscientiousness	780
Emotional stability	671

Table 6.1. Number of training items per trait.

Note: We have chosen to perform a content analysis by personality trait instead of facet analysis. Facet analysis requires identifying and collecting enough structured and high-quality training items per facet to develop an efficient classification model. The performance of the model can be significantly impacted with fewer than 40 examples per class (Fyffe, Lee & Kaplan, 2023). Given the complexity of creating a training set by facet, we have opted for an intermediate analysis by trait for now.

On the other hand, based on the results obtained by Fyffe, Lee, and Kaplan (2023), we have chosen to use DeBERTa. DeBERTa - "Decoding-enhanced BERT with disentangled attention" - is a transformer-based natural language processing (NLP) model (He, Liu, Tao & Chen, 2021). DeBERTa is an enhancement of the Bidirectional Encoder Representations from Transformers (BERT) model, which is one of the most successful NLP models to date. DeBERTa uses a transformer architecture similar to BERT, but it features several enhancements and innovations to improve performance on different NLP tasks, including improved decoding, deinterlaced attention, multi-task adaptation, and model compression. To date, DeBERTa has shown outstanding performance on a wide range of NLP tasks, including text classification.

6.1.4. Results

This classification model has learned to effectively classify personality items into six traits from a large corpus of assessments. In other words, based on the content of an item, this model can determine the personality trait most related to that item. The goal is to use this model to classify SWIPE items into the six personality traits they are intended to measure to determine what they actually measure and assign them a trait. The traits assigned by the model will constitute the "reference" trait, as it is the trait to which the item is most representative. These labels will then be compared to the traits initially assigned to each

SWIPE item, which constitute the "prediction". If we have classified the items in the same trait as our model, it means they are representative of the trait they measure. If we have classified the items in a different trait than our model, it means they are not representative of the trait they are supposed to measure, or they are more representative of another trait. It's important to note that although the classification results of the model are taken as a reference source due to its performance, automated models for classifying personality items being often more efficient than human judges (83% accuracy for DeBERTa, 71% for a human judge, see Fyffe, Lee and Kaplan, 2023), it would be naive not to keep this margin of error in mind, and to maintain a critical eye on the results.

To assess the content validity of the items for each trait, four indicators are measured:

- Accuracy: measures the proportion of correct predictions compared to all the predictions made. It is therefore the ability to correctly predict positive and negative observations. It is calculated by dividing the total number of correct predictions by the total number of predictions made. The accuracy varies from 0 to 1;

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives}$$

- Precision: measures the proportion of positive predictions that are correct among all predictions made, regardless of whether they are actually positive or negative. In other words, precision measures our ability to correctly identify positive cases. It is calculated by dividing the number of true positive predictions by the total number of positive predictions (both true and false). The precision varies from 0 to 1;

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

- Recall: measures the proportion of true positive examples that are correctly predicted among all positive examples. In other words, recall measures the ability to find all positive observations. It is calculated by dividing the number of correct positive predictions by the total number of actual positive examples. The recall varies from 0 to 1;

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

- F1-score: a combined measure of precision and recall. It is the harmonic mean of precision and recall. The F1-score can be considered as the overall indicator of efficiency. The F1-score ranges from 0 to 1, where a value of 1 indicates optimal performance in terms of precision and recall;

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

From a general point of view, there are no universally good or bad scores for each of these measures. The scores depend on the context and the specific requirements of the classification problem. For example, in some applications such as financial fraud detection, high accuracy may be crucial to minimise the number of false positives, even though it may reduce recall and miss some cases of fraud. In other applications, such as email spam detection, high recall may be more important to ensure that all spam messages are identified, although it may increase the number of false positives. However, it is generally accepted that, for each of these indicators, and in particular for the F1-score, a score above or equal to 0.9 is considered excellent, a score between 0.8 and 0.9 is good, a score between 0.7 and 0.8 is satisfactory, a score between 0.5 and 0.7 is passable, and a score below or equal to 0.5 is considered very insufficient.

The results obtained by SWIPE items are presented in Table 6.2.

Traits	Accuracy	Precision	Recall	F1-score
Extraversion	.75	.87	.75	**.81**
Agreeableness	.89	1	.89	**.94**
Humility	1	.70	1	**.82**
Openness	.96	1	.96	**.98**
Conscientiousness	.93	.96	.93	**.95**
Emotional stability	.96	.92	.96	**.94**
	.92	*.91*	*.92*	*.91*

Table 6.2. Performance by personality trait.

6.1.5. Results interpretation

If this type of analysis is new and is certainly the first of its kind in the study of content validity for the development of a new personality assessment, the presented results provide valuable insights into the quality of SWIPE items and their representativeness of the personality traits they are intended to measure:

- For Agreeableness, the results are excellent, with a precision of 1, a recall of .89, and an F1-score of .94. This means that most of the items belonging to Agreeableness (as defined by our model) were correctly classified during the development of SWIPE, and all the items classified in this trait by SWIPE are also classified by our reference model;
- For Openness, the results are excellent, with a precision of 1, a recall of .96, and an F1-score of .98. Most of the items belonging to Openness (as defined by our model) were correctly classified during the development of SWIPE, and all the items classified in this trait by SWIPE are also classified by our reference model;
- For Conscientiousness, the results are excellent, with a precision of .96, a recall of .93, and an F1-score of .95. Most of the items belonging to Conscientiousness (as defined by our model) were correctly classified during the development of SWIPE, and all items classified in this trait by SWIPE are also classified by our reference model;
- For Emotional stability, the results are excellent, with a precision of .92, a recall of .96, and an F1-score of .94. Most of the items belonging to Emotional stability (as defined by our model) were correctly classified during the development of SWIPE, and all items classified in this trait by SWIPE are also classified by our reference model;

- For Extraversion, we observe a high precision (.87) but a lower recall (.75), indicating that whilst most of the items classified as Extraversion by SWIPE are correct, some items that should have been classified as Extraversion (as defined by our model) were instead classified into another trait during the development of SWIPE;

- For Humility, the precision is lower (.70) and the recall is much higher (1), meaning that all items classified by our reference model as representative of Humility were correctly assigned to this trait by SWIPE, but that SWIPE also assigned some items to Humility that our reference model did not classify as such, leading to a lower precision score.

If we look at the results presented, they are excellent for four traits, but more attention needs to be paid to the traits of Extraversion (F1-score = .81) and Humility (F1-score = .82), which are slightly set back. A thorough analysis of the items that were the subject of divergent classification helps to find a conceptual explanation. It appears that most of the items initially classified in Humility during the development of SWIPE are either identified by our model as more representative of Extraversion or Agreeableness. This classification pattern probably arises from the natural and demonstrated links that exist between these personality traits. The "assertiveness" and "sociability" facets, which belong to Extraversion, show negative correlations with Humility (Lee, Ashton & De Vries, 2022; Ludeke, Bainbridge, Liu, Zhao, Smillie & Zettler, 2019), whilst Agreeableness is positively correlated (Lee, Ashton & De Vries, 2022). Similarly, Humility is strongly linked to the Dark Triad (Howard & Van Zandt, 2020), which is itself highly correlated with assertiveness (Kaufman, Yaden, Hyde & Tsukayama, 2019). Finally, our own studies on a sample of participants who completed the BFI-2 show significant correlations between several facets related to these personality traits, including assertiveness ~ modesty ($r = -.36$; $p < 2.2e\text{-}16$), respectfulness ~ modesty ($r = .39$; $p < 2.2e\text{-}16$), and respectfulness ~ sincerity ($r = -.35$; $p < 2.2e\text{-}16$). These findings are also confirmed by the inter-dimension correlations obtained in the literature (Soto & John, 2017). Therefore, given the links between these facets and personality traits, it is not inconsistent to see items classified differently between the language model and the SWIPE model.

Figure 6.1. Classification clusters of SWIPE items by the NLP model.

This spatial representation allows for a better visualisation of the few classification errors mentioned earlier. Some items that were initially classified in Humility (green dots on the graph) were instead classified into Extraversion (dark blue dots on the graph) or Agreeableness (red dots on the graph). This cluster analysis thus confirms the close links between these three personality traits (Lee, Ashton & De Vries, 2022; Ludeke, Bainbridge, Liu, Zhao, Smillie & Zettler, 2019; Soto & John, 2017). These links are also demonstrated by the spatial proximity of the three clusters on the graph.

Conversely, Openness (light blue), Conscientiousness (pink) and Emotional Stability (brown) are represented by clusters which are more independent.

Also, instead of questioning the results and quality of the SWIPE items, this conclusion highlights the need to consider potential biases inherent in the trained classification model. This model was trained with the assumption that a response choice could only measure a single personality trait, which is known as multiclass classification. However, many publishers of personality assessments are now using "blended items" that measure different facets of personality (Schwaba, Rhemtulla, Hopwood & Bleidorn, 2020), as is the case with SWIPE. Therefore, it is necessary to focus on multilabel classification techniques, which involve assigning multiple labels to a single observation (Fyffe, Lee et Kaplan, 2023) - this means that each observation, or each response choice in our specific case, may belong to more than one category simultaneously. Multilabel classification is generally more complex than multiclass classification because it requires a more detailed analysis of each observation to determine the different labels that correspond to it. Multilabel classification algorithms can be more computationally expensive, but they are often more flexible and suitable for certain tasks. We are currently conducting studies on this topic, and we will update this guide with our findings when they become available.

6.1.6. Comparison with other assessments

To go further in understanding the results and the quality of the SWIPE items, we can also compare the indicators obtained for SWIPE with those obtained for other personality assessments. These analyses will be added to this manual when they become available.

6.1.7. Conclusion

The results demonstrate excellent content validity of the SWIPE scales at the trait level. Indeed, the indicators measured show that the SWIPE items are representative of the personality traits they measure. The results are excellent for 4 of the 6 traits measured (Agreeableness, Openness, Conscientiousness, and Emotional Stability), and good for the other two (Extraversion and Humility). However, it should be mentioned that the results concerning the last two traits mentioned are negatively impacted by a conceptual overlap between the two traits, and by the fact that our model was trained by multiclass classification. Although the results remain good, it is important to consider these limitations. In summary, the content validity results presented here attest to the theoretical, conceptual, and semantic soundness of SWIPE, and demonstrate that its content is representative of the Big Five and the added scale of humility.

6.2. Construct validity

6.2.1. Introduction

Construct validity refers to whether an assessment instrument measures the intended theoretical construct and not something else. It is closely related to other aspects of validity, as any evidence of validity contributes to understanding the construct validity of a test. The importance of construct validity lies in the fact that it influences the interpretation of test scores. If a test claims to measure a specific personality facet, it is crucial to ensure that it actually measures that facet. Otherwise, any interpretation of the scores would be incorrect and could lead to biased decisions. However, construct validity is not limited to simply looking at whether the assessment is measuring a specific facet. It involves a comprehensive investigation to determine whether the interpretations of the test results are consistent with the theoretical and observational terms that define the construct (Cronbach & Meehl, 1955).

There is no single method for determining construct validity, but rather different methods and approaches must be combined. In order to assess the construct validity of SWIPE, we have utilised four complementary methods: item-dimension saturation and inter-dimension correlation (Thurstone, 1947; Bollen, 1989; McDonald, 2013), the Root Mean Square Error of Approximation (RMSEA), and the presentation of distribution parameters (Fisher, 1912, 1920, 1921, 1922).

- **Item-dimension saturation** refers to the correlation between an item and the total score of the dimension or factor to which it belongs. In other words, if an item is designed to measure a particular facet, it should be closely associated with other items that measure that facet. Thus, the higher the correlation between an item and the dimension, the more strongly the item is related to that dimension and therefore more valid. For item-dimension saturation, a value of .40 or higher is generally considered satisfactory and adequate. This suggests that the item measures the dimension it is supposed to measure (Campbell & Fiske, 1959; Nunnally, 1978; Hair, Black, Babin, & Anderson, 2010). Saturation below .40 may be acceptable if supported by theoretical justification;

- **Inter-dimension correlation** assesses the relationship between scores of different factors or dimensions measured by a test. If two dimensions are expected to be distinct and independent, then they should have weakly correlated scores. On the other hand, if the dimensions are closely related or overlapping, the scores should be more strongly correlated. There is no universal threshold for inter-dimension correlation. It is generally desirable for the dimensions to be relatively independent, although there may be some moderate correlations between the dimensions that are justified by the underlying theoretical model. If the correlations between the dimensions do not match theoretical expectations, this may indicate a construct validity issue. It is, therefore, necessary to be able to compare these correlations with those of the foundational and reference theoretical construct (the Big Five Inventory-2 in the case of the SWIPE assessment);

- The **RMSEA**, or Root Mean Square Error of Approximation, measures the difference between the observed data and the fitted data of the model, corrected for the number of free parameters of the model. The RMSEA assesses the absolute fit of the model by comparing the unexplained variance in the data with the expected unexplained variance in the population given the model. Generally, an RMSEA < .05 indicates a good fit of the model to the data (Steiger & Lind, 1980; Browne & Cudeck, 1993);

- **The distribution parameters** correspond to the statistical characteristics of the distribution of scores in the assessments. These parameters include measures such as the mean, standard deviation, skewness, and kurtosis, which provide information about the shape, centre, and variability of the distribution. They make it possible to identify atypical scores, explore individual differences in the distribution of scores, and to better interpret the results. For example, high levels of skewness or kurtosis may indicate non-normal distributions, which could affect the interpretation of the results and the use of certain statistical tests. Therefore, it is important to examine distribution parameters in addition to other aspects of validity when assessing the quality of an assessment.

6.2.2. Item-dimension saturation

The latest SWIPE item-dimension saturation studies were conducted in April 2023 (N = 4,457) on the main items measuring each facet. The table below presents the results of this analysis: (1) the saturation varies from .30 ≤ r ≤ .68, (2) the average saturation per facet varies from .46 ≤ r ≤ .58. These conclusions attest to satisfactory and adequate item-dimension saturation.

Traits	Facets	i1	i2	i3	i4	i5	i6	i7	i8	Mean
EXTRAVERSION	Assertiveness	.68	.64	.52	.49	.48	.46	.53	.41	**.53**
	Energy level	.58	.54	.42	.45	.58	.52	.55	.47	**.52**
	Sociability	.51	.51	.66	.58	.48	.49	.64	.41	**.54**
AGREEABLENESS	Compassion	.59	.30	.42	.63	.50	.38	.53	.58	**.49**
	Respectfulness	.41	.57	.56	.35	.41	.55	.51	.31	**.46**
	Trust	.53	.48	.54	.56	.35	.49	.44	.44	**.48**
HUMILITY	Greed avoidance	.57	.51	.51	.50	.50	.40	.63	.41	**.50**
	Modesty	.46	.52	.53	.49	.44	.51	.54	.47	**.49**
	Sincerity	.54	.44	.51	.56	.52	.38	.45	.42	**.48**
OPENNESS	Aesthetic sensitivity	.35	.68	.63	.52	.63	.30	.60	.51	**.53**
	Creative imagination	.61	.53	.50	.54	.51	.63	.62	.57	**.56**
	Intellectual curiosity	.52	.46	.55	.52	.49	.41	.65	.48	**.51**
CONSCIENTIOUSNESS	Organisation	.58	.51	.58	.47	.52	.45	.57	.62	**.54**
	Productiveness	.53	.47	.60	.58	.61	.45	.41	.42	**.51**
	Responsibility	.50	.49	.41	.48	.49	.47	.46	.44	**.47**
EMOTIONAL STABILITY	Anxiety	.52	.42	.47	.61	.64	.63	.44	.40	**.52**
	Depression	.54	.59	.67	.49	.57	.55	.58	.61	**.57**
	Emotional volatility	.44	.48	.44	.59	.48	.50	.59	.49	**.50**

Table 6.4. Item-dimension saturation for SWIPE facets.

6.2.3. Inter-dimension correlation

SWIPE's latest inter-dimension correlation studies were conducted in April 2023 (N=4,457). To study the dynamics of these inter-dimension correlations with regard to the underlying theoretical model and to validate their consistency, several analyses were proposed. These include (1) a study of inter-dimension correlations of SWIPE, presented in table 6.5, (2) a study of the inter-dimension correlations of the BFI-2, presented in table 6.6, (3) an analysis of consistency between the two correlation matrices using the rank correlation of Spearman or ρ of Spearman, (4) an analysis of the effect size using Cohen's q, presented in table 6.7, and (5) a theoretical explanatory review of the links between facets, presented in table 6.8.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		.47	.56	-.12	-.29	.04	.01	.24	.32	-.04	.32	-.02	-.32	-.61	-.11	-.37	-.42	-.18
2			.60	.17	.07	.37	-.03	.12	.14	.05	.56	.08	-.05	-.18	.17	-.54	-.74	-.36
3				.17	-.09	.31	-.01	.14	.22	-.15	.23	-.17	-.09	-.31	.09	-.32	-.41	-.02
4					.47	.51	.18	.09	.05	-.07	.07	.06	.38	.41	.43	.02	-.09	.09
5						.37	.09	-.05	-.13	.16	.11	.35	.30	.47	.31	-.08	-.14	-.26
6							.10	.10	.07	-.17	.09	-.09	.26	.23	.29	-.35	-.39	-.17
7								.52	.47	-.05	-.02	.02	.14	.13	.13	.05	.01	.11
8									.49	-.19	.08	-.04	.07	-.04	.09	-.04	-.07	.09
9										-.09	-.01	-.08	-.04	-.19	.01	-.08	-.10	.05
10											.38	.60	-.14	.06	.14	-.02	-.09	-.20
11												.52	-.07	-.09	.20	-.24	-.43	-.31
12													-.02	.17	.25	-.01	-.13	-.27
13														.56	.37	.07	.05	.05
14															.39	.20	.17	.04
15																.04	-.10	.00
16																	.80	.67
17																		.61
18																		

Table 6.5. Inter-dimension correlation for SWIPE facets.

Overall, facets are weakly correlated, supporting an acceptable level of consistency. The comparison with the inter-dimension correlations of the BFI-2 allows us to go further in our understanding of these correlations, by analysing them with regard to the underlying theoretical model.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		.49	.48	-.02	-.07	.11	.03	.30	.17	.09	.31	.17	-.21	-.36	-.06	-.38	-.45	-.25
2			.50	.27	.22	.33	.09	.31	.09	.22	.53	.28	.01	-.07	.19	-.39	-.61	-.28
3				.15	-.01	.21	.04	.16	.05	.00	.21	.04	-.05	-.27	.03	-.22	-.33	-.05
4					.43	.38	.19	.15	.12	.13	.21	.23	.28	.35	.38	.03	-.16	.02
5						.33	.10	.07	.02	.25	.29	.46	.19	.39	.36	-.12	-.25	-.27
6							.10	.13	.05	.05	.18	.12	.19	.16	.27	-.29	-.35	-.21
7								.41	.45	.00	.02	.05	.02	.00	.05	.06	.00	.07
8									.40	-.03	.16	.09	.01	-.08	.03	-.15	-.20	-.07
9										-.07	-.02	.02	.02	-.10	-.01	.01	-.02	.02
10											.51	.48	.01	.17	.23	-.06	-.22	-.21
11												.52	.11	.15	.29	-.28	-.48	-.33
12													.11	.25	.32	-.21	-.37	-.42

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
13														.42	.33	-.01	-.01	-.04
14															.41	.07	.00	-.09
15																-.05	-.19	-.10
16																	.66	.62
17																		.58
18																		

Table 6.6. Inter-dimension correlation for BFI-2.

In order to ensure consistency between the two matrices, we use Spearman's ρ as a measure of non-parametric correlation between two variables. Unlike the Pearson correlation, which measures the linear relationship between two continuous variables, the Spearman correlation assesses the monotonic relationship between two variables. Spearman's correlation uses the ranks of the observations of each variable instead of their actual values to calculate the correlation. Observations are ranked in ascending or descending order according to their value, and corresponding ranks are assigned. Spearman's correlation ranges from -1 to 1, where a value of 1 indicates a perfect positive monotonic relationship, a value of -1 indicates a perfect negative monotonic relationship and a value of 0 indicates no monotonic relationship between the variables. Using Spearman's ρ is an appropriate choice to assess the consistency between the inter-dimension correlations of SWIPE and the BFI-2 matrices.

The results indicate a value of $\rho = .66$ ($p < 2.2e\text{-}16$), demonstrating the consistency between the two matrices. In other words, the inter-dimension correlations observed in SWIPE are also reflected in the underlying theoretical construct, indicating that they are inherent to the facets and concepts being measured.

To go further in understanding the convergences and divergences between the two matrices, we also propose an analysis of the size effects with Cohen's q. Cohen's coefficient is a measure of the effect of the size of a difference between two groups in a statistical study. Cohen's coefficient ranges from -1 to 1, where 0 indicates no difference between groups, 1 indicates maximum difference, and -1 indicates maximum difference the other way. In general, a value $q \approx .0$ indicates no difference, a value $q \approx .3$ corresponds to a small difference, $q \approx .5$ corresponds to a medium difference, and $q \approx .8$ corresponds to a strong difference. Also, to ensure the consistency of the inter-dimension correlations between the two matrices, we seek to obtain values of q as close as possible to 0. The results of this analysis are presented in the table below.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		-.02	.11	-.10	-.23	-.08	-.03	-.07	.15	-.14	.01	-.19	-.12	-.33	-.05	.01	.03	.08
2			.14	-.10	-.15	.05	-.12	-.19	.05	-.17	.03	-.21	-.06	-.11	-.02	-.19	-.24	-.09
3				.02	-.08	.10	-.06	-.02	.17	-.15	.02	-.21	-.04	-.04	.06	-.10	-.09	.04
4					.05	.16	-.01	-.06	-.06	-.21	-.14	-.18	.11	.07	.07	-.02	.07	.06
5						.04	-.01	-.12	-.16	-.10	-.19	-.13	.12	.09	-.07	.04	.12	.01
6							.00	-.02	.02	-.22	-.09	-.22	.08	.07	.01	-.06	-.04	.04
7								.15	.03	-.05	-.04	-.03	.13	.14	.08	-.02	.01	.04
8									.11	-.16	-.09	-.13	.06	.04	.07	.11	.13	.16
9										-.02	.01	-.09	-.06	-.09	.02	-.09	-.08	.03
10											-.16	.17	-.16	-.11	-.09	.05	.14	.02

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
11												.01	-.18	-.24	-.10	.04	.06	.03
12													-.14	-.08	-.08	.20	.26	.17
13														.18	.04	.08	.07	.09
14															-.03	.14	.17	.13
15																.09	.09	.10
16																	.30	.09
17																		.04
18																		

Table 6.7. Analysis of Cohen's q coefficients.

In conclusion, the inter-dimension correlations in SWIPE are relatively low and correspond to theoretical expectations. Indeed, there is no significant difference between the inter-dimension correlations highlighted in SWIPE and those in the BFI-2, indicating that the two matrices are equivalent. Furthermore, the strongest inter-dimension correlations relate to facets whose links have repeatedly been identified and justified in the scientific literature:

Facet n°1	Facet n°2	Reference
Assertiveness	Energy level	Soto & John, 2017; DeYoung, Quilty & Peterson, 2007; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Sociability	Assertiveness	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Sociability	Energy level	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Respectfulness	Compassion	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Trust	Compassion	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Creative imagination	Aesthetic sensitivity	Soto & John, 2017; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Intellectual curiosity	Aesthetic sensitivity	Soto & John, 2017; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Intellectual curiosity	Creative imagination	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Productiveness	Energy level	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Responsibility	Organisation	Soto & John, 2017; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Responsibility	Productiveness	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Modesty	Assertiveness	Lee, Ashton, & de Vries, 2022; Ludeke, Bainbridge, Liu, Zhao, Smillie & Zettler, 2019.
Modesty	Greed avoidance	Denissen, Soto, Geenen, John & van Aken, 2022.
Anxiety	Energy level	Halama, Kohút, Soto & John, 2020; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.

Depression	Energy level	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Depression	Anxiety	Soto & John, 2017; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Emotional volatility	Anxiety	Soto & John, 2017; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.
Emotional volatility	Depression	Soto & John, 2017; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Føllesdal & Soto, 2022; Halama, Kohút, Soto & John, 2020; Gallardo-Pujol, Rouco, Cortijos-Bernabeu, Oceja, Soto & John, 2021; Vedel, Wellnitz, Ludeke, Soto, John & Andersen, 2021.

Table 6.8. Theoretical review of the inter-dimension correlations.

6.2.4. RMSEA

The latest RMSEA studies for SWIPE were conducted in April 2023 (N=4,457). For each facet, the RMSEA index was less than .05 (mean RMSEA = .01), indicating a good fit of the model to the data. This suggests that the model has a good ability to explain the relationships between the measured variables and that the differences between the observed data and the data predicted by the model are small. Overall, these results provide additional evidence of construct validity for SWIPE.

Facets	RMSEA	Facets	RMSEA
Assertiveness	.010	Aesthetic sensitivity	.015
Energy level	.006	Creative imagination	.016
Sociability	.011	Intellectual curiosity	.015
Compassion	.008	Organization	.017
Respectfulness	.001	Productiveness	.013
Trust	.010	Responsibility	> .001
Greed avoidance	.018	Anxiety	> .001
Modesty	.015	Depression	.011
Sincerity	.018	Emotional volatility	.011

Table 6.9. RMSEA for each facet.

6.2.5. Distribution parameters

The distribution of scores obtained from a personality assessment is an essential aspect of the assessment's construct validity. The way scores are distributed for each personality facet can provide vital information about how the test measures that facet and how the scores are interpreted. Our analysis of distribution parameters focuses on six primary parameters that are necessary:

- the **mean,** which is an indicator of the central tendency of the scores in the distribution;
- the **median**, which is a measure of central tendency that represents the value that divides a distribution in half, with 50% of the scores above and 50% below the median. Unlike the mean, the median is less sensitive to extreme scores and is a more robust measure of central tendency;
- the **standard deviation**, which is a measure of the dispersion of scores around the mean. It is calculated by taking the square root of the variance of the scores;
- the **range** of scores;
- **skewness**, which is a measure of the symmetry of a distribution. It is calculated by comparing the frequency of scores to the left and right of the mean. If the distribution is perfectly symmetric, the skewness is zero. If the distribution is skewed to the left, the skewness is negative. If the distribution is skewed to the right, the skewness is positive;
- **kurtosis,** which is a measure of the degree of peakedness or flatness of a distribution compared to a normal distribution. A normal distribution has a kurtosis of 0. If a distribution is more peaked than a normal distribution, its kurtosis value is positive, and if it is less peaked, its kurtosis value is negative.

Expectations for distribution parameters depend on the context and the measurement instrument used. However, in general, here is what is expected for "good" distribution parameters:

- the mean should be close to the median value: this indicates that the distribution is symmetric. If the mean is significantly different from the median, this may indicate an asymmetry in the distribution;
- the standard deviation should be reasonable and large enough to capture individual differences in the measured dimension, but not so large as to dilute the differences between individuals. In general, one would expect the standard deviation to be around 2 for the 10-point personality scale;
- the range should capture the variation in the measured dimension, but not be so large as to dilute the differences between individuals. In general, the scale is expected to be between 4 and 6;
- the asymmetry (skewness) should be close to 0 (symmetrical distribution). If the skewness is significantly different from 0, this may indicate an asymmetry in the distribution;
- the kurtosis should be close to 0 (normal distribution). If the kurtosis is significantly different from 0, the distribution is either flatter or more peaked.

It is important to note that these expectations may vary depending on the context and the measurement instrument used. For example, for some personality assessments, it may be normal to have a skewed distribution or a larger range. In this sense, it is necessary to put into perspective the results presented below for SWIPE with those generally obtained in the scientific literature with the BFI. Also, several studies have demonstrated slightly more negative kurtosis for the BFI (Plaisant, Courtois, Réveillère, Mendelsohn & John, 2009; Rammstedt, 2007; DeYoung, Carey, Krueger & Ross, 2016). For example, a study by Plaisant, Courtois, Réveillère, Mendelsohn, and John (2009), relating to the validation of BFI in the French language, showed a slightly negative kurtosis – between -.53 and .56. It should also be noted that slight asymmetries and kurtosis are not unusual in measures of personality facets, and do not call into question the validity and reliability of the BFI-2. The distributions of personality scores are often slightly asymmetrical or with kurtosis coefficients different from zero.

SWIPE's latest distribution parameter studies were conducted in April 2023 (N=4,457).

Facets	Mean	Median	Standard deviation	Range	Skewness	Kurtosis
Assertiveness	5.75	6	2.19	3.65	.1	-.8
Energy level	5.42	5	2.15	4.19	-.18	-.13
Sociability	5.57	6	2.05	4.39	.22	-.15
Compassion	5.51	5	1.93	4.66	.02	-.21
Respectfulness	5.54	6	2	4.50	.18	-.1
Trust	5.92	5	2.34	3.42	.05	-1.1
Greed avoidance	5.67	5	2.16	4.17	.27	-.36
Modesty	5.62	5	2.13	4.23	.32	-.14
Sincerity	5.81	5	2.41	3.73	.41	-.6
Aesthetic sensitivity	5.56	6	2.06	4.37	.23	-.09
Creative imagination	5.69	5	2.2	4.09	.39	-.28
Intellectual curiosity	5.52	5	1.96	4.59	.11	-.14
Organization	5.48	6	1.98	4.55	-.09	-.08
Productiveness	5.49	5	1.95	4.62	-.07	-.15

Responsibility	5.45	5	1.94	4.12	.15	-.33
Anxiety	5.67	5	2.2	4.09	.36	-.22
Depression	5.32	6	1.89	4.76	.32	.9
Emotional volatility	5.6	6	2.05	4.39	.19	-.15
	5.59	*5.39*	*2.09*	*4.25*	*.17*	*-.23*

Table 6.10. SWIPE's distribution parameters.

The distribution parameters are thus consistent with the expected standards, and also with the scientific literature relating to the underlying theoretical model, the BFI-2. The normality of the distributions can be assessed by indices such as the similarity of means and medians, as well as the skewness and kurtosis coefficients, which are close to 0.

6.2.6. Conclusion

Several results have demonstrated the construct validity of SWIPE, including: (1) meeting the required standards for inter-dimension saturations, (2) having weak inter-dimension correlations that largely converge with those inherent in the constructs measured and are supported by the scientific literature, (3) having RMSEA indices for each facet that largely respect appropriate thresholds, and (4) having good distribution parameters that reflect those theoretically expected. These findings suggest that SWIPE accurately measures the facets it claims to assess.

6.3. Convergent validity

6.3.1. Introduction

Convergent validity is a measure of how similar the scores of a personality test are to scores from other tests or measures that assess the same personality dimension or factor. This allows us to verify whether a personality test accurately measures what it is intended to measure. Specifically, convergent validity is determined by the correlation between the scores of a test and those of other measures or tests that assess the same facet of personality. A strong correlation between the scores indicates that the scales are measuring the same construct, which strengthens the validity of the test.

It should be noted that there is no "official" threshold for judging the quality of convergence between two measures. Additionally, the appropriate threshold depends on the specific context in which the assessment is used and the characteristics of the target population. Furthermore, convergent validity must be assessed in conjunction with other measures of validity to have a complete assessment of the quality of the personality test. However, several authors and researchers have offered some suggestions or satisfaction thresholds: (1) a correlation of .7 or more between an assessment and other measures that assess the same dimension is an indicator of very strong convergent validity, according to Campbell and Fiske (1959); (2) a correlation of .6 is recommended as a threshold of validity by Worthington and Whittaker (2006); (3) a correlation of .5 or more is considered good convergent validity by Bagozzi and Yi (1988) and by Revelle and Condon (2015); (4) a correlation of .4 is considered acceptable by Nunnally and Bernstein (1994). In short, although there is no clear consensus or golden rule (Marsh, Hau & Wen, 2004) on the exact value to use as the threshold of convergent validity, it is recommended to aim for correlations of .5 or higher to support good convergent validity of a personality assessment.

6.3.2. Convergent validity with BFI-2

We are studying the convergent validity of SWIPE with the BFI-2 (Lignier, Petot, Canada, Oliveira, Nicolas, Courtois, John, Plaisant & Soto, 2022). The most recent studies assessing the convergent validity of SWIPE with the BFI-2 were conducted in April 2023 (N=4,457).

SWIPE facets	BFI-2 facets	*r*
Assertiveness	Assertiveness	.77
Energy level	Energy level	.71
Sociability	Sociability	.72
Compassion	Compassion	.58
Respectfulness	Respectfulness	.55
Trust	Trust	.70
Greed avoidance	Greed avoidance	.62
Modesty	Modesty	.61
Sincerity	Sincerity	.58
Aesthetic sensitivity	Aesthetic sensitivity	.66
Creative imagination	Creative imagination	.70
Intellectual curiosity	Intellectual curiosity	.64
Organization	Organization	.66
Productiveness	Productiveness	.70
Responsibility	Responsibility	.55
Anxiety	Anxiety	.76
Depression	Depression	.72
Emotional volatility	Emotional volatility	.72

Table 6.11. Convergent validity (r) between SWIPE and BFI-2 facets.

6.3.3. Conclusion

The presented analyses demonstrate good convergent validity of SWIPE with the BFI-2, with correlations ranging from .55 to .77. Also, the correlations of nine facets exceed the threshold of .7 proposed by Campbell and Fiske (1959), and all correlations exceed the threshold of .5 proposed by others. As a result, we can conclude that the relationships between SWIPE and the BFI-2 are strong enough to validate a base of similar constructs.

6.4. Predictive validity

The predictive validity of a personality assessment measures its ability to predict a target variable, such as job performance or employee turnover. The question is whether the results of the personality assessment can accurately predict future work-related outcomes. Evidence of predictive validity is particularly useful when one wants to make inferences about an individual's future performance or behaviour based on their scores on the assessment. Currently, studies on the predictive validity of SWIPE are ongoing and will be soon added to this technical guide.

6.5. Conclusion

Validation of a personality assessment is crucial to ensure the accuracy of the results. In this study, we examined the content validity, construct validity, and convergent validity of SWIPE. Our analyses show that SWIPE covers the theoretical constructs it is designed to measure, the assessment is well-structured and exhibits good measurement homogeneity, and there is a strong correlation between SWIPE and the BFI-2, confirming the similarity of the constructs measured. Overall, the results obtained meet the most demanding psychometric standards and demonstrate the validity of SWIPE: it does measure the presented personality facets. However, further investigation into the psychometric qualities of SWIPE requires an examination of its reliability. In this sense, an assessment must be both valid and reliable to be used in HR decisions (such as recruitment, mobility, etc.). A valid but not reliable assessment would indicate that the test measures what it should measure, but the individual scores are inconsistent. On the contrary, a valid and reliable assessment ensures that the assessment consistently measures what it is supposed to measure. In other words, it hits the bullseye consistently. The evidence of reliability is presented in the next chapter.

Summary of validity

The purpose of validating an assessment is to confirm that it actually measures what it is designed to measure, and to determine the accuracy of the results obtained from it. Validation studies typically focus on content validity, construct validity, convergent validity, and predictive validity.

.91 Mean F1-Score. The results demonstrate excellent content validity of the SWIPE scales at the trait level.

.01 The mean RMSEA indicates a good fit of the model to the data and provides evidence of the construct validity of SWIPE.

.70 Average correlation with the BFI-2 scales, which demonstrates the convergent validity of SWIPE, and support the measure of similar constructs.

7. Reliability

How can you determine if the results of an assessment are reliable? How can you ensure that the assessment produces consistent results when asking the same questions to the same person at different times? The answers to these questions can be obtained through the study of reliability. Whilst validity provides information on an assessment's ability to measure what it intends to measure, reliability measures whether the measurement is consistent and reliable every time the same assessment is completed by the same person. In short, the reliability of an assessment measures its consistency or stability over time and aims to determine if an assessment produces similar results when asking the same questions to the same person at different times or to similar people. Therefore, the objective of reliability is to ensure that the obtained results are dependable and accurate. The reliability of a assessment can be evaluated in two different and complementary ways:

- **Internal consistency**, which is a statistical measure used to assess the reliability of a psychometric test. It evaluates the homogeneity or similarity of different test items that are intended to measure the same psychological dimension. In other words, internal consistency assesses whether multiple items that are designed to measure the same thing produce similar scores;

- **Test-retest reliability**, which is a method used to assess the reliability of a measurement by measuring the same variable at two different points in time. This approach enables the assessment of the temporal stability of the measurement and the estimation of the proportion of total variance attributable to measurement error. Test-retest reliability is frequently utilised in longitudinal studies or to evaluate the stability of a test over a specific period.

7.1. Internal consistency

The concept of internal consistency was introduced by psychologist Lee Cronbach in the 1950s. He proposed Cronbach's alpha as a measure of the reliability of an assessment, which calculates the average correlation between the different items. Cronbach's alpha has since been widely used as a measure of internal consistency in psychometric testing. However, whilst Cronbach's alpha has gained popularity due to its ease of calculation and interpretation, it has several limitations in assessing the reliability of more modern assessments. Firstly, it is difficult to obtain high internal consistencies in forced-choice assessments, as this format distorts the internal consistency of instruments (Brown & Maydeu-Olivares, 2013). Secondly, Cronbach's alpha tends to underestimate reliability (Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019). Thirdly, it is more suitable for one-dimensional scales, where each item measures only one facet (Cortina, 1993). Finally, its use is strongly influenced by the number of items, the number of orthogonal dimensions, and the mean of the correlations between the items (Cortina, 1993). Therefore, its use is increasingly criticised and not recommended.

To address these limitations, several authors recommend the use of another indicator: McDonald's Omega, which was introduced by J.B. McDonald in 1970 as an alternative to Cronbach's alpha. McDonald, an American psychologist, developed this measure of reliability based on a factorial approach. Omega has two advantages in particular: (1) it takes into account the strength of the association between the items and a construct, and (2) it takes into account the link between the items and the measurement error. Since its inception, McDonald's Omega has been widely used and validated in numerous studies. For example, a study by Revelle and Zinbarg (2009) showed that Omega was the best reliability index among 12 in total. Other studies have since confirmed these results, solidifying McDonald's Omega as the most appropriate coefficient for accurately judging the reliability of personality scales (Kelley & Pornprasertmanit, 2015; Trizano-Hermosilla & Alvarado, 2016; Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019). It is now often recommended as a replacement for Cronbach's alpha.

Since then, other methods have emerged to overcome the difficulties encountered by Cronbach's alpha (Green and Yang, 2009; Osburn, 2000; Revelle and Zinbarg, 2009; Sijtsma, 2009; Trizano-Hermosilla and Alvarado, 2016). In particular, the lambda2, lambda4, and lambda6 indicators have gained attention (see definition table below). These measures are based on the early work of Guttman (Guttman, 1945), who identified six types of coefficients (lambda1 to lambda6) and showed that each was a lower bound for the true reliability, defined as the ratio of the variance from the actual score to the variance of the observed score (Guttman, 1945; Callender and Osburn, 1979). As synthesised by Bourque, Doucet, LeBlanc, Dupuis, and Nadeau (2019), "lambda1 greatly underestimates the true fidelity and is not used as a fidelity estimator but as an intermediate step for other calculations" (p. 82), (2) lambda3 is mathematically equivalent to Cronbach's alpha, (3) "lambda-5, on the other hand, is efficient when there is a high covariance between one item and the others, which, in turn, do not have a high covariance between them, which is undesirable in the case of a psychometric scale" (p. 83). Among these lambda indicators, we favour those with the greatest empirical support in estimating real reliability, namely lambda2, lambda4, and lambda6. These indicators are further defined in the table below.

Method	Description	Reference
lambda2	Lambda2 is a lower bound of reliability that equals the true reliability if the test items are tau-equivalent. Lambda2 is interesting because it always provides a lower bound that is as good as alpha, but can be significantly better in other cases. Lambda2 is always higher than lambda1 and is greater than or equal to lambda3 (i.e., Cronbach's alpha) if there is independence between item errors.	Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019; Momirović, 1996; Malkewitz, Schwall, Meesters & Hardt, 2023; Guttman, 1945; Callender & Osburn, 1977; Callender & Osburn, 1979; Sijtsma, 2009; Thompson, Green & Yang, 2010; Osburn, 2000; van der Ark, van der Palm & Sijtsma, 2011; Cho, 2022; Revelle, 1979; Tang & Chui, 2012; Hunt & Bentler, 2015; Benton, 2013; Berge & Socan, 2004.
lambda4	Lambda4 is calculated by dividing the assessment into two random halves, using the split-half method. Then, the covariance between the scores obtained on each half of the assessment is calculated, and the variance of the total assessment score is also calculated. Lambda4 is generally considered taking the split which maximises reliability. It, therefore, represents bisection coefficient.	
lambda6	Lambda-6 reflects the proportion of the total variance of an item that is explained by the linear regression of that item on all other items in the scale. It is also known as the squared multiple correlation coefficient and is a measure of the degree to which an item is related to the overall construct being measured.	

Table 7.1. Description of lambda measures.

It is important to note that these indicators are not interchangeable, and their choice will depend on the objectives of the study and the characteristics of the measurement scale. However, several studies have shown that: (1) Cronbach's alpha is one of the least effective indicators, (2) Cronbach's alpha and lambda2 systematically and significantly underestimate reliability, (3) the best index would be Omega in the case where there are few items, (4) and lambda6 in all other cases (Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019). Also, although lambda4 is a reliability coefficient that remains interesting in terms of ease of understanding and is less likely to underestimate reliability, as Cronbach's alpha can, it may tend to overestimate reliability if there are a large number of items or if the sample size is small (Benton, 2013; Berge & Socan, 2004). However, given the nature of SWIPE, consisting of a limited number of items, and the fairly large sample used for our studies, this risk is minimised. In short, although all these indicators are proposed for the study of SWIPE reliability, the most consistent ones remain the Omega, lambda4, and lambda6.

In addition to the aforementioned analyses, we propose a study of measurement errors (Kim & Feldt, 2010) for SWIPE. Measurement error refers to the random variation in personality measurement that can arise due to measurement errors or external factors that affect test results. Several factors can contribute to this error, including individual differences in test comprehension, scoring or coding errors, variations in the test-takers' mindset or mood, or measurement method errors. Measurement error can adversely affect the reliability and validity of personality test results by producing scores that do not accurately reflect the test-takers' personality facets. Therefore, minimising measurement error is crucial. To study measurement errors, the following analysis is proposed in this chapter:

- the presentation of information and measurement error curves for each facet;
- empirical reliability (empirical_rxx), which is calculated based on the data obtained during the administration of a test to a sample of people, and reflects the reliability of the assessment as measured from empirical data;
- marginal reliability (marginal_rxx), which is estimated based on a statistical model that considers the structure of test scores and measurement errors. It provides a theoretical estimate of the reliability.

The acceptability thresholds for each of these indicators have varied over time, depending on factors such as the type of assessment, number of items, or distribution of participant responses. Nevertheless, typical values include: (1) .6-.7 for Cronbach's alpha (Nunnally, 1978), (2) .7 for McDonald's Omega (McDonald, 1999), (3) .6 for the lambda indicators (Callender & Osburn, 1979), (4) .6-.7 for both empirical_rxx and marginal_rxx (Chalmers, 2012).

Finally, an investigation of the inter-item correlation is conducted to assess the degree of correlation among items that measure a particular personality facet. Specifically, the inter-item correlation refers to the average correlation among each item, which helps determine the assessment's internal consistency and the extent to which the items measure the same construct. Unlike Cronbach's α, the average inter-item correlation is considered a simpler indicator of a scale's internal consistency as it minimises the effects of the total number of items. Typically, an ideal level of homogeneity is achieved when the inter-item correlation for a facet falls between .15 and .40 (Piedmont & Hyland, 1993). Values below .1 suggest that the items are too different and measure distinct constructs, whilst a correlation exceeding .4 indicates that the items are too similar and redundant. Overall, the acceptable threshold ranges from .15 to .50 (Clark & Watson, 1995).

7.1.1. Cronbach's alpha

The most recent studies on Cronbach's alpha for SWIPE were conducted in April 2023 (N=4,457). With the exception of five dimensions that have α coefficients between .64 and .70, all other α coefficients are greater than .70, indicating adequate reliability results for SWIPE. However, it is also important to consider the short, forced-choice, and multidimensional structure of SWIPE, which limit the relevance of Cronbach's alpha as an indicator in this context.

Facets	α	Facets	α
Assertiveness	.75	Aesthetic sensitivity	.70
Energy level	.77	Creative imagination	.76
Sociability	.75	Intellectual curiosity	.66
Compassion	.70	Organisation	.74
Respectfulness	.66	Productiveness	.71
Trust	.65	Responsibility	.73
Greed avoidance	.68	Anxiety	.79
Modesty	.73	Depression	.81
Sincerity	.64	Emotional volatility	.71

Table 7.2. Cronbach's alpha α for each facet.

7.1.2. McDonald's Omega

The latest studies on McDonald's Omega for SWIPE were conducted in April 2023 (N=4,457). The Omega coefficients are all greater than .70, indicating satisfactory reliability results for SWIPE. Additionally, given the relevance of McDonald's Omega in the context of reliability analyses, these results appear more appropriate for evaluating the stability and internal consistency of SWIPE. These findings provide evidence for the consistency of SWIPE scales.

Facets	ω	Facets	ω
Assertiveness	.78	Aesthetic sensitivity	.74
Energy level	.81	Creative imagination	.79
Sociability	.79	Intellectual curiosity	.70
Compassion	.72	Organisation	.78
Respectfulness	.72	Productiveness	.74
Trust	.71	Responsibility	.76
Greed avoidance	.73	Anxiety	.81
Modesty	.76	Depression	.84
Sincerity	.70	Emotional volatility	.74

Table 7.3. McDonald's Omega ω for each facet.

7.1.3. Lambda measures

The latest lambda indicator studies for SWIPE were conducted in April 2023 (N=4,457) and pertain to lambda2, lambda4, and lambda6. Overall, all values are above .70, with the trust, sincerity, and intellectual curiosity facets having the lowest values, but still remaining very close to .70. These results demonstrate the overall reliability of SWIPE.

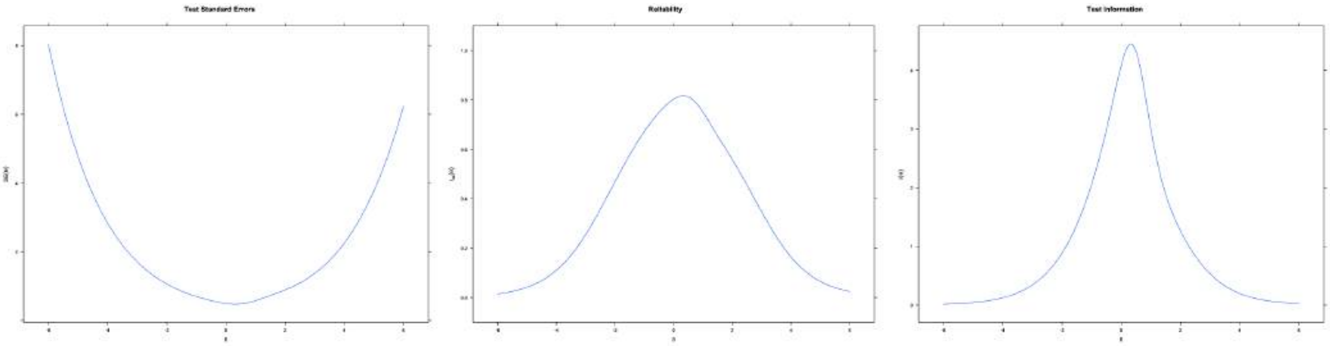
Facets	lambda2	lambda4	lambda6
Assertiveness	.75	.79	.75
Energy level	.78	.84	.78
Sociability	.76	.80	.75
Compassion	.70	.73	.68
Respectfulness	.68	.72	.67
Trust	.66	.72	.65
Greed avoidance	.70	.74	.68
Modesty	.73	.77	.72
Sincerity	.65	.71	.63
Aesthetic sensitivity	.70	.73	.68
Creative imagination	.76	.80	.75
Intellectual curiosity	.66	.72	.64
Organization	.75	.77	.73
Productiveness	.71	.76	.70
Responsibility	.74	.78	.73
Anxiety	.79	.83	.79
Depression	.82	.84	.81
Emotional volatility	.71	.75	.70
	.72	*.77*	*.71*

Table 7.4. Lambda2, lambda4 and lambda6 for each facet.

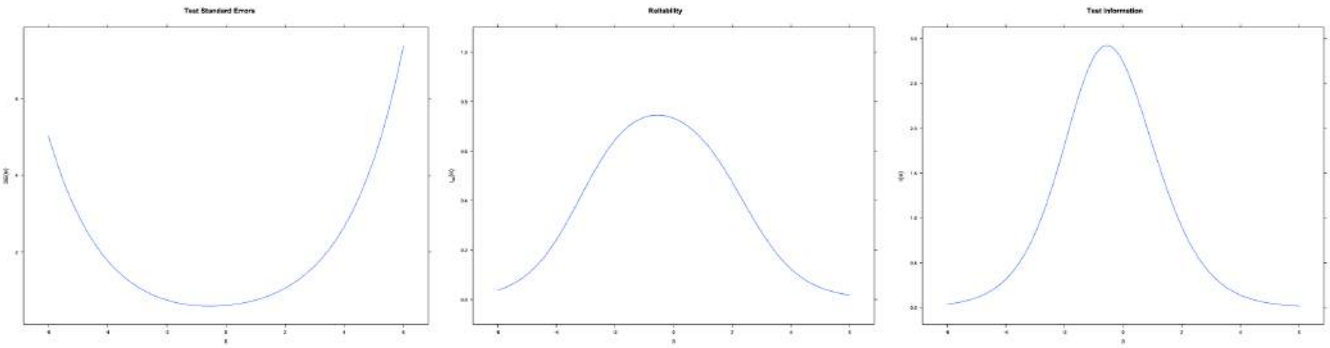
7.1.4. Measurement errors and IRT reliability

The latest studies on measurement errors, empirical and marginal reliability for SWIPE were conducted in April 2023 (N = 4,457). For each facet, the following information is presented: (1) measurement error, which is represented by the test standard errors SE(θ); (2) reliability, which is represented by rxx(θ); and (3) the information curve (test information).

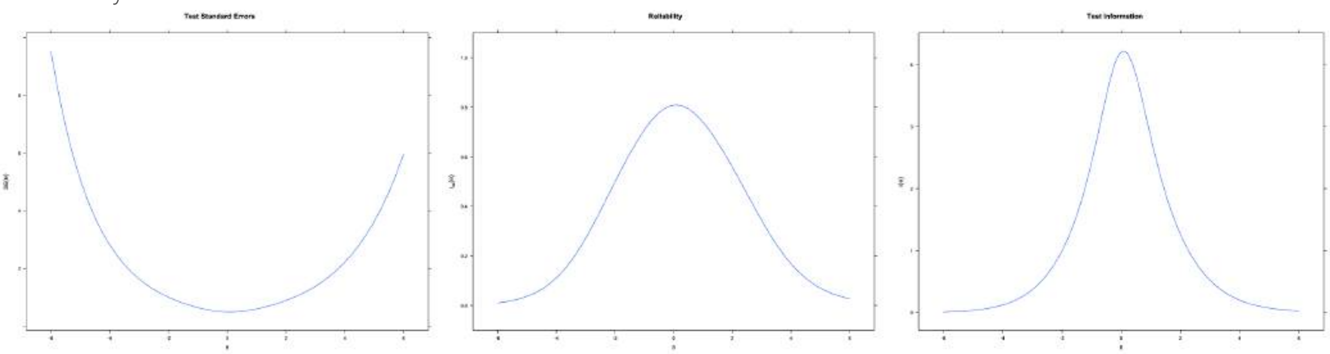
Assertiveness.



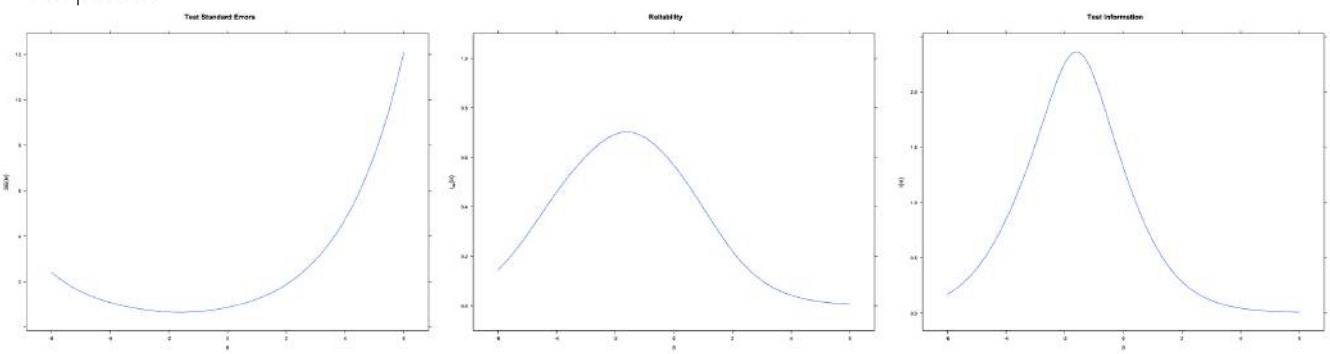
Energy level.



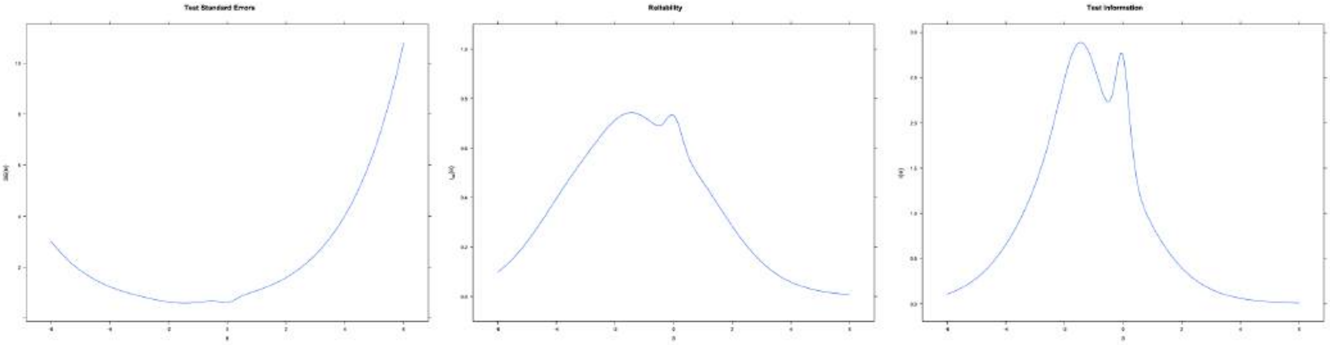
Sociability.



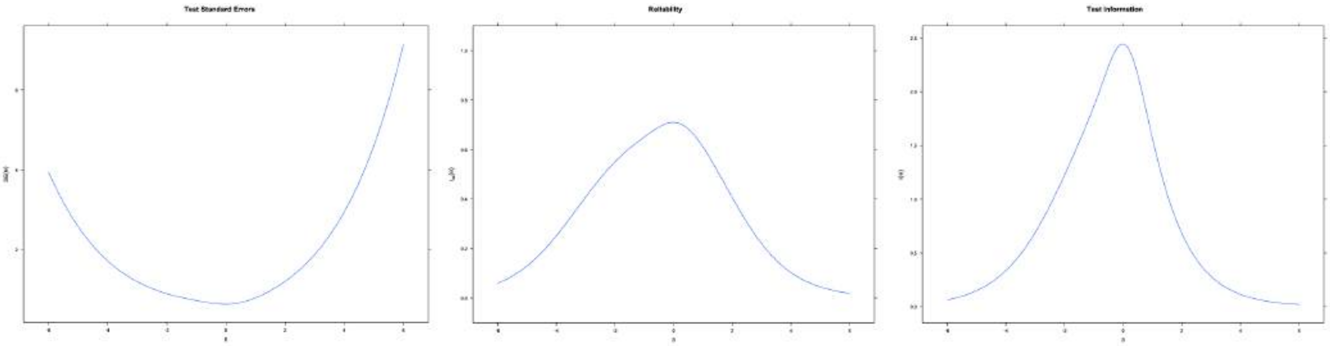
Compassion.



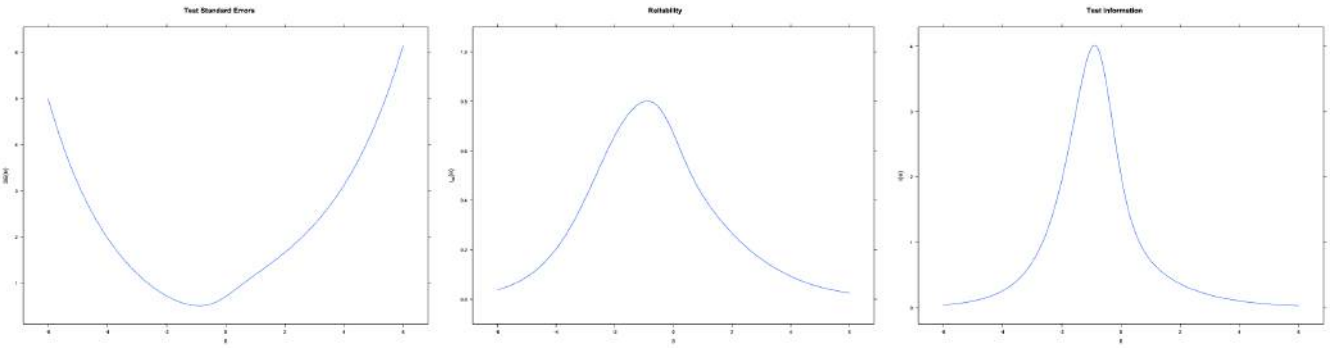
Respectfulness.



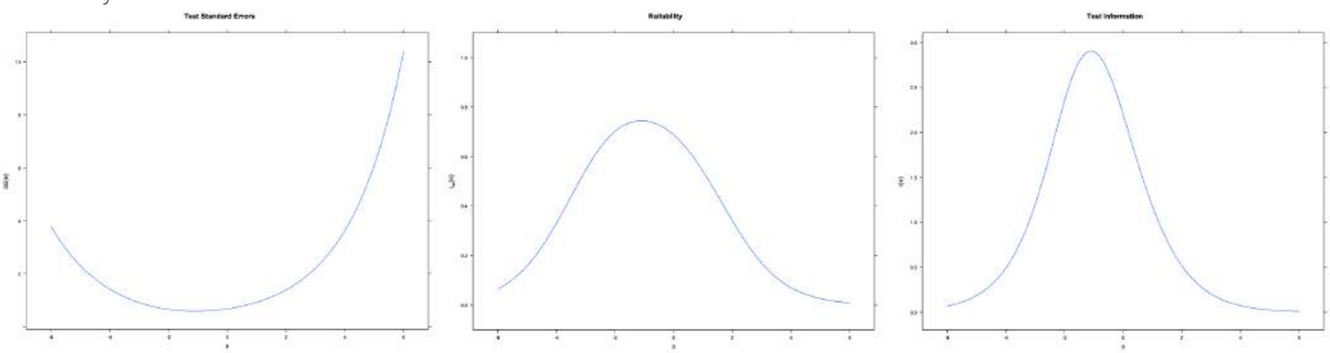
Trust.



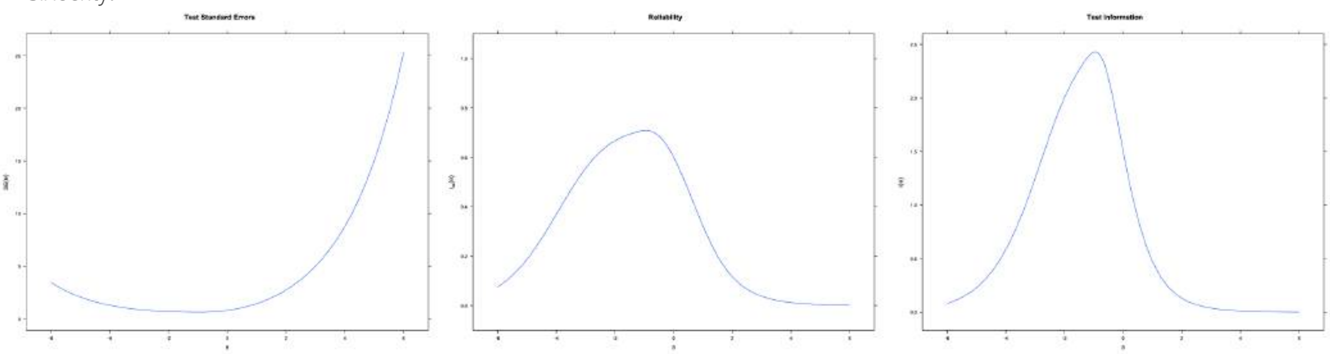
Greed avoidance.



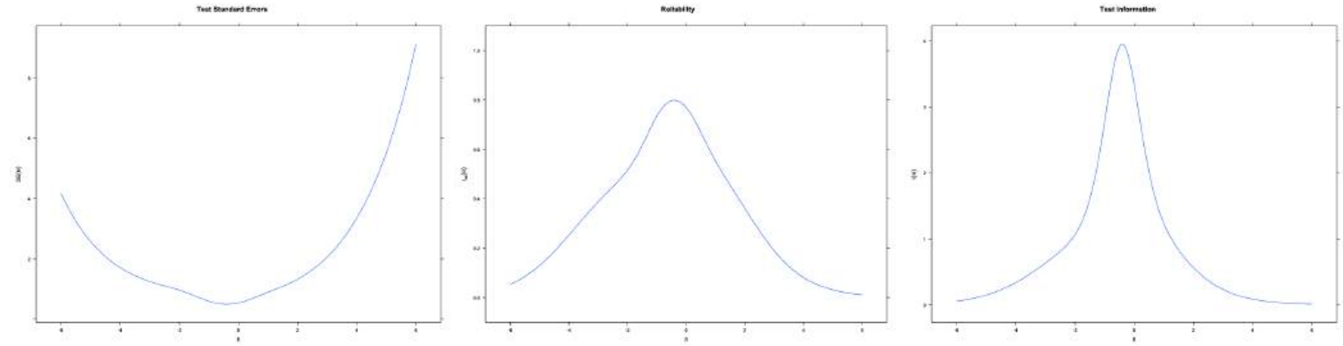
Modesty.



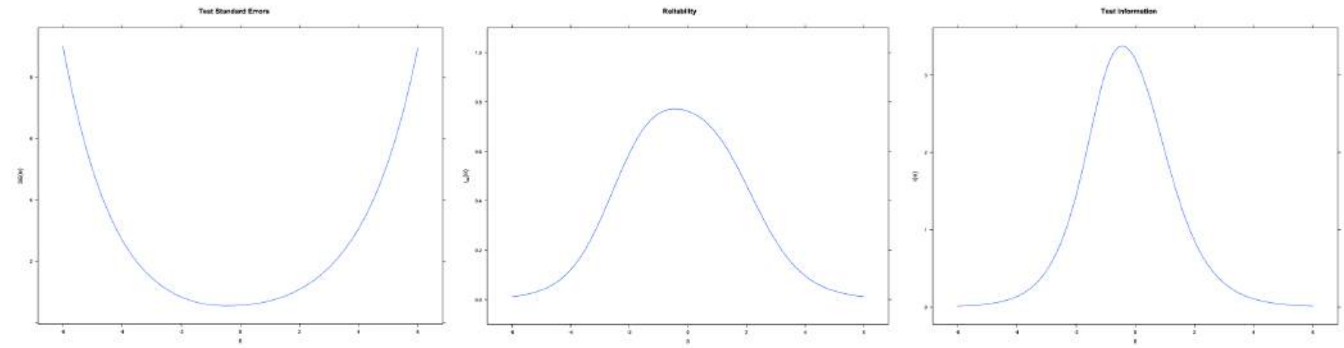
Sincerity.



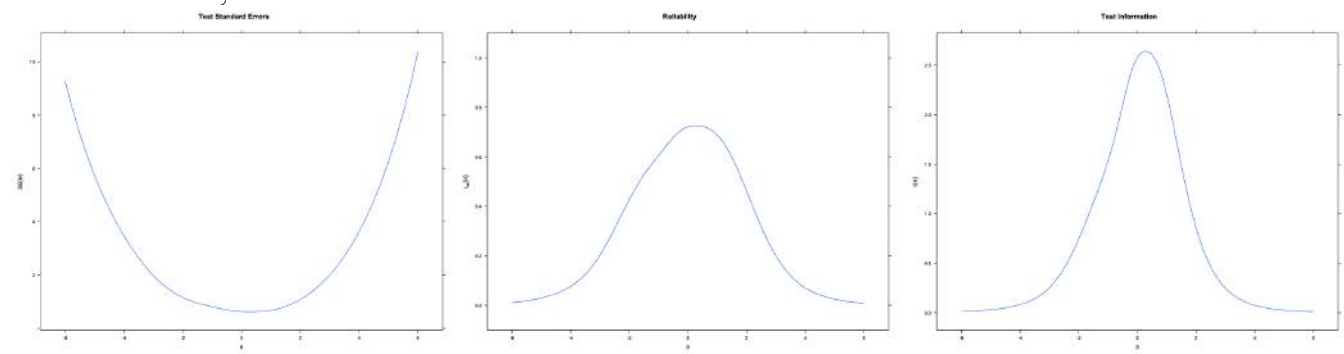
Aesthetic sensitivity.



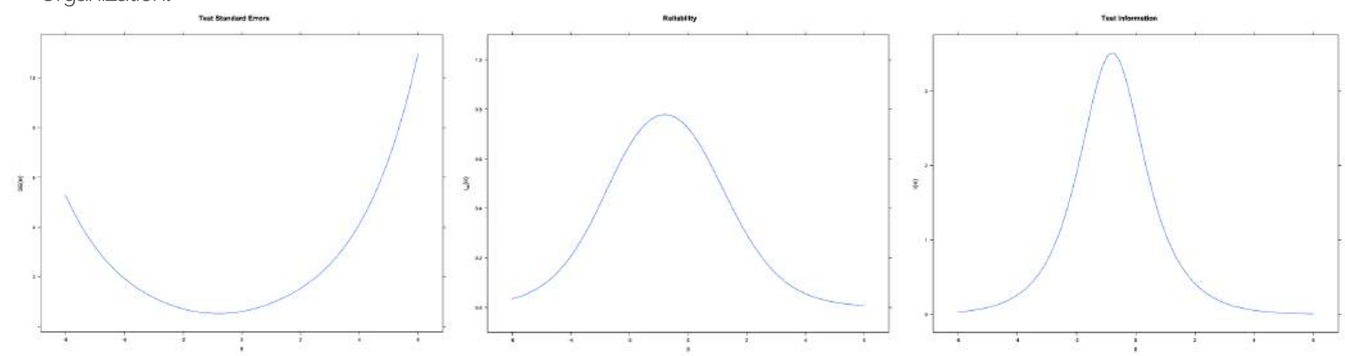
Creative imagination.



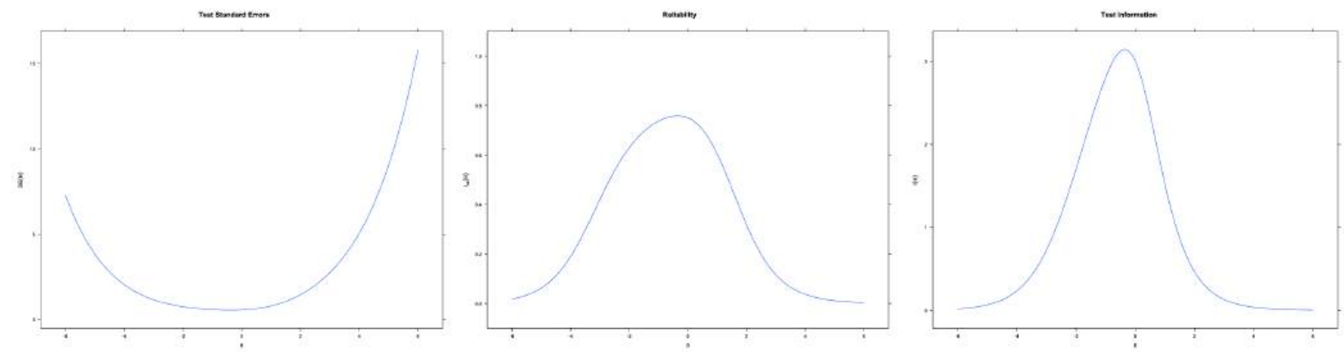
Intellectual curiosity.



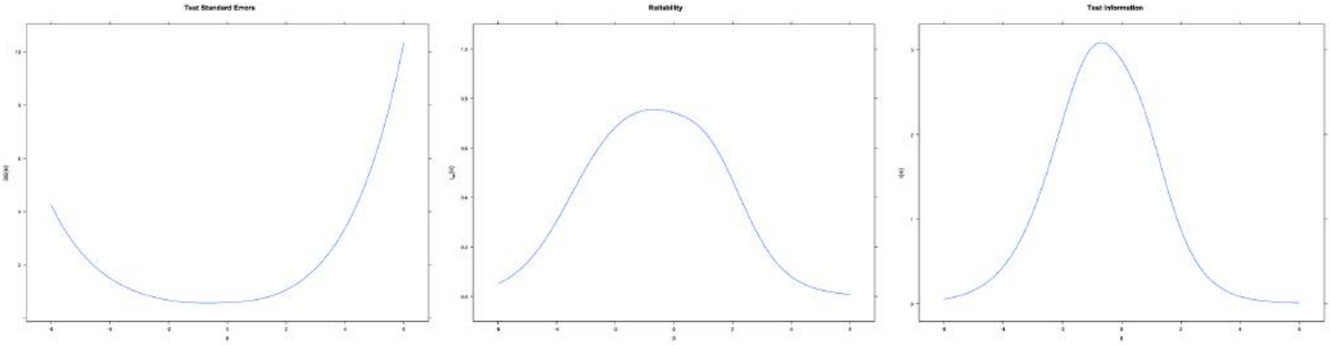
Organization.



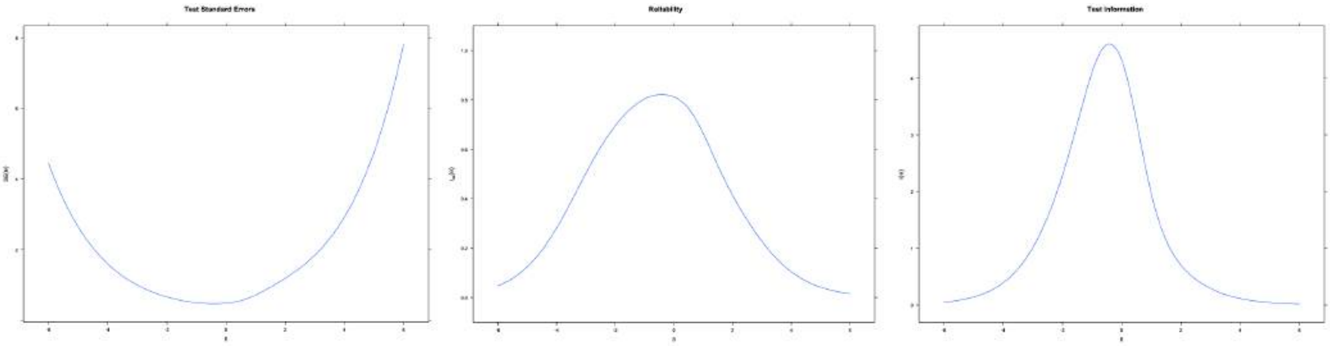
Productiveness.



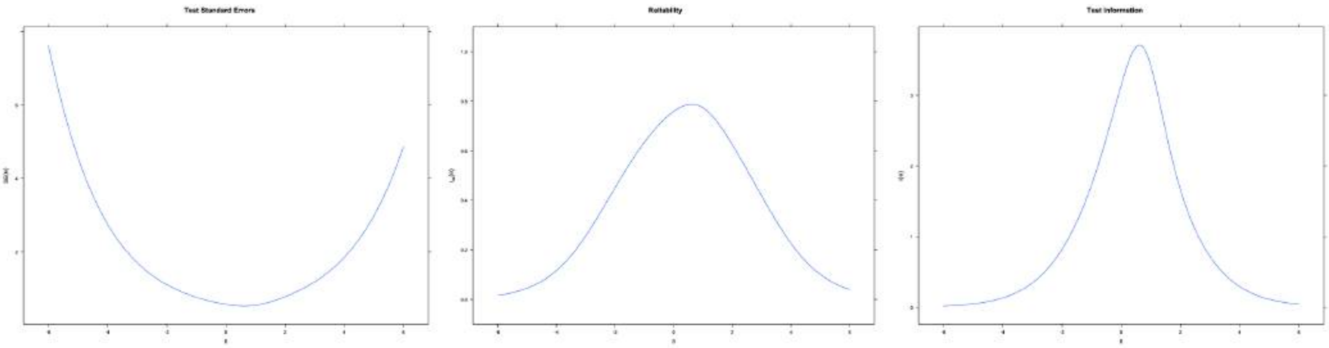
Responsibility.



Anxiety.

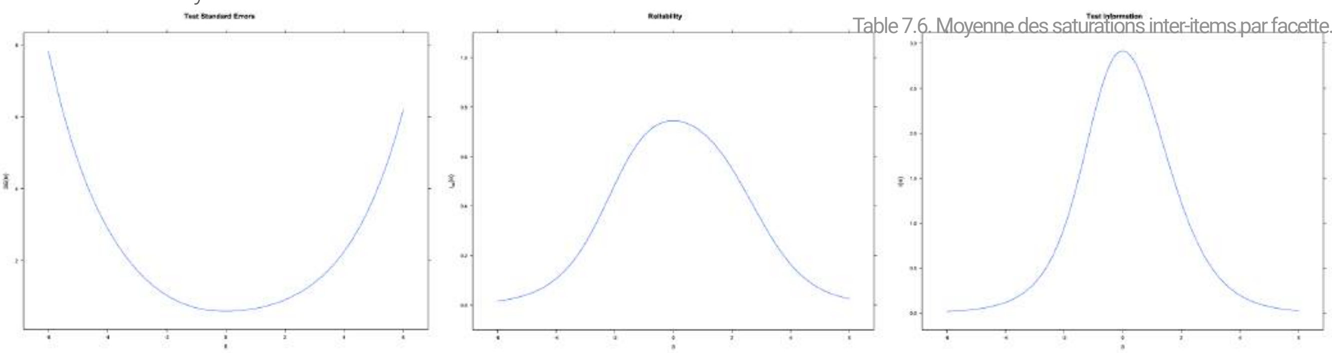


Depression.



Emotional stability.

Table 7.6. Moyenne des saturations inter-items par facette.



The tables below present the values of empirical_rxx and marginal_rxx for each personality facet in SWIPE. The recommended threshold values for both indicators are between .6 and .7 (Chalmers, 2012). With the exception of the Sincerity facet (empirical_rxx = .57 and marginal_rxx = .53), all other facets meet these recommended thresholds, providing further evidence for the reliability of SWIPE.

Facets	empirical_rxx	marginal_rxx	Facets	empirical_rxx	marginal_rxx
Assertiveness	.74	.72	Aesthetic sensitivity	.70	.67
Energy level	.73	.70	Creative imagination	.73	.71
Sociability	.75	.73	Intellectual curiosity	.66	.65
Compassion	.60	.54	Organization	.70	.66
Respectfulness	.63	.62	Productiveness	.70	.70
Trust	.65	.64	Responsibility	.72	.71
Greed avoidance	.64	.61	Anxiety	.77	.74
Modesty	.70	.64	Depression	.76	.70
Sincerity	.57	.53	Emotional volatility	.70	.70
				.70	*.67*

Table 7.5. Empirical_rxx and marginal_rxx for each facet.

7.1.5. Inter-item correlation

The latest inter-item correlation studies for SWIPE were conducted in April 2023 (N=4,457). The average inter-item correlation coefficients are all between .15 and .50, indicating an optimal level of homogeneity for each facet of the assessment (Piedmont & Hyland, 1993; Briggs & Cheek, 1986; Clark & Watson, 1995). Therefore, the items measuring a specific facet are well connected to each other, but not to the extent that they become redundant. Each item provides unique and specific information.

Facets	MIC	Facets	MIC
Assertiveness	.21	Aesthetic sensitivity	.21
Energy level	.22	Creative imagination	.26
Sociability	.23	Intellectual curiosity	.18
Compassion	.21	Organization	.25
Respectfulness	.18	Productiveness	.20
Trust	.16	Responsibility	.20
Greed avoidance	.21	Anxiety	.23
Modesty	.20	Depression	.30
Sincerity	.18	Emotional volatility	.20

Table 7.6. Mean of inter-item correlation for each facet.

7.1.6. Conclusion

The presented internal consistency analysis results indicate that SWIPE is a reliable assessment that meets the required standards of quality and measurement fidelity. The high Cronbach's alpha and McDonald's Omega coefficients show that the assessment items are strongly related to each other and consistently measure the same personality dimension. These conclusions are particularly noteworthy considering the underestimation biases associated with Cronbach's alpha. The McDonald's Omega coefficients are more adapted to the measurement context and the SWIPE scales. The lambda indicators, especially lambda4 and lambda6, also confirm and demonstrate the internal consistency of SWIPE and its overall reliability.

7.2. Test-retest reliability

The test-retest reliability measures the temporal consistency of an assessment or measurement scale by administering the same assessment to a group of participants at two different times with a time interval between the two. The correlation between the two results is then calculated to determine the reliability of the test. A high correlation indicates that the participants' scores are stable over time, indicating the assessment's reliability. Test-retest reliability is crucial for personality assessments to ensure that the results are consistent and reliable over the long term (Spearman, 1904; Thorndike, 1918; Guilford, 1936; Anastasi, 1954; Cronbach, 1951). Several recent studies have analysed the test-retest reliability of the BFI and BFI-2, demonstrating high test-retest reliability for the five personality traits, with correlations ranging from .63 to .86 (Zhang et al., 2022; Courtois et al., 2018, 2020; Seybert & Becker, 2019; Gnambs, 2016).

Test-retest reliability studies are typically conducted at 3, 6, and 9-month intervals. Also, SWIPE being a new assessment, the time intervals are currently too short to ask people to retake the assessment. Test-retest reliability studies linked to SWIPE will, therefore, be carried out within appropriate time intervals to conduct a reliable study: in July 2023 (t+3 months), in October 2023 (t+6 months), and in January 2024 (t+9 months). The results of these studies will be added to this technical manual as soon as possible.

Summary of reliability

Reliability refers to the extent to which a measurement or assessment produces consistent results over time and across different situations. It aims to determine whether an assessment consistently measures what it is supposed to measure and produces similar results each time it is administered to the same group of people.

.72 Mean Cronbach's Alpha, showing adequate reliability results from SWIPE, despite the underestimation of this indicator.

.76 Average McDonald's Omega, demonstrating the strong consistency of SWIPE's scales. The indicator is typically recommended and better.

.77 Average lambda4. Lambda4 measure is a more appropriate indicator for the nature of SWIPE, and it confirms its reliability.

8. Sensitivity

Sensitivity, also called discrimination, refers to the ability of an assessment to distinguish between people with a high level on a facet and people with a low level. It, therefore, reflects the ability of the assessment to identify the uniqueness of each individual. A sensitive personality assessment can identify the subtle differences between people, and help to understand their behaviours with more precision and discrimination. Sensitivity thus refers to the ability of the assessment to correctly identify people who possess the characteristic being measured and to avoid false positives (people who do not possess the characteristic, but who are identified as such by the assessment).

The first attempts to measure discrimination were based on cumulative scales (Guttman, 1944; Walker, 1931; Loevinger, 1948; Loevinger, 1953, cited by Hankins, 2008). However, Ferguson was one of the first to propose conceptualising discrimination in the form of a coefficient. In this sense, if there is a maximum number of possible differences in a sample, the discrimination coefficient corresponds to the ratio between the number of differences actually observed and this maximum number of differences. This coefficient called the delta δ of Ferguson (Ferguson, 1949; Kline, 2000), is thus the ratio between the differences observed between people and the number of maximum possible differences. It is intended to be a direct and non-parametric index of the degree of distinction made by an instrument between individuals. If no difference is observed, then $\delta = 0$. If all possible discriminations are observed, then $\delta = 1$. Generally, a normal distribution should have excellent discrimination, where $\delta \geqslant .9$ (Ferguson, 1949). Weaker discriminations are expected for leptokurtic distributions (because these distributions fail to discriminate around the mean) and skewed distributions (because these fail to discriminate at one end of the distribution). Demonstrating excellent discrimination between the scales of an assessment requires a $\delta \geqslant .9$ for each scale. For example, in a recent adaptation study of the BFI-2 in Russian, Kalugin, Shchebetenko, Mishkevich, Soto, and John (2021) showed that all scales had strong discriminations.

The latest sensitivity studies for SWIPE, with Ferguson's δ, were conducted in April 2023 (N = 4,457). The δ coefficients for all measurement scales were found to be greater than .9 (mean $\delta = .95$), indicating excellent discrimination. This means that the assessment is able to accurately distinguish individual differences in personality among the people who took the assessment and that it is sensitive to variations in the measured personality facets. The results are presented in Table 8.1.

Facets	δ	Facets	δ
Assertiveness	.96	Aesthetic sensitivity	.96
Energy level	.96	Creative imagination	.96
Sociability	.96	Intellectual curiosity	.96
Compassion	.93	Organization	.96
Respectfulness	.96	Productiveness	.96
Trust	.96	Responsibility	.90
Greed avoidance	.96	Anxiety	.96
Modesty	.94	Depression	.97
Sincerity	.91	Emotional volatility	.96

Table 8.1. Delta of Ferguson (δ) for each SWIPE facet.

How to properly read the results: Ferguson's δ is the ratio between the differences observed between people and the number of maximum possible differences. A δ of 0 means that no discriminations are made by the scale, whilst a δ of 1 means all possible discriminations are made. For example, for the "Sociability" facet, $\delta = .96$, which means that 96% of all possible discriminations are made by the "Sociability" scale.

9. Fairness

Fairness in the context of a personality assessment refers to the extent to which it is designed to be fair and unbiased for all individuals, regardless of their origin, gender, sexual orientation, race, or culture. In other words, a fair assessment should be objective and impartial towards all individuals who take it, without any bias or discrimination against any particular group. Our teams take every measure to ensure the fairness of our assessments and predictive analyses, and we ensure that the use of our algorithms in decision-making processes does not lead to discrimination through any unforeseen algorithmic biases. Additionally, in the development of our assessments, equity studies focus on two areas: (1) ensuring the accessibility of the assessment, and (2) ensuring equity in the results of the assessment.

9.1. SWIPE accessibility

The user experience and accessibility of the solution are important priorities for AssessFirst. We, therefore, care about offering an assessment process and a results interface that are easy to use and understand. The efforts we deploy are what make AssessFirst an essential player when it comes to user experience today: the experience we offer is fluid, transparent, and above all, it addresses everyone, regardless of age, profession, degree, or mastery of digital tools, etc. The Google Reviews from our candidates, available here, are a testimony to this. The actions implemented by AssessFirst to ensure and improve the accessibility of SWIPE include:

- **Professional nature of the content**: SWIPE and its results were specifically developed to be relevant in a professional context. The dimensions assessed were selected for their relevance to professional efficiency. The conclusions drawn from the use of AssessFirst are limited to this specific context;
- **Language level**: AssessFirst relies on a Localisation team made up of psychologists and experts in linguistic management, in order to provide textual content that is understandable and accessible to all, in all languages (15 languages currently available). We work with native-language translators to create and validate all of our content;
- **Psychometric properties**: SWIPE has been developed to meet the most demanding psychometric standards in terms of validity, reliability, and sensitivity;
- **Fairness by design**: We aim to create our assessments using neutral content that does not reference any cultural or social codes. Additionally, in SWIPE, the amount of text to read has been significantly reduced, with 65% less text than in SHAPE, for example. This effort to decrease the volume of text enhances the accessibility of SWIPE to individuals with reading disorders;
- **Text-to-speech**: We have developed our own text-to-speech tool to automatically read assessments. This feature provides access to a vocal assistant that reads the items, reinforcing accessibility for people with visual disabilities;
- **Management of contrasts**: AssessFirst implements actions to allow personalisation of contrast and display settings of web content to make it easier to read for users with visual impairments;
- **Customer integration**: Many partnerships have been implemented with target customers who offer the solution to populations who may have difficulty accessing the tool (e.g. users with disabilities, users with little access to employment, young populations, populations who lack digital literacy, and populations without professional experience). Regular discussions with these partners and users allow us to continuously improve the solution in order to better meet their needs.

These initiatives drive the accessibility of AssessFirst solutions, reflecting our unwavering commitment to ensuring that all users can benefit from their results, gain a better understanding of their unique strengths, and develop their talents to their full potential. Our ongoing partnerships have yielded results that position

AssessFirst as a leading innovator in the HR Tech industry, setting new standards for inclusivity and diversity. To date, our solutions have impacted the lives of over 5,000,000 individuals, each given the opportunity to be recognised for their true worth as human beings, rather than being judged by factors such as their academic or professional background, age, or gender. We remain dedicated to these objectives, and these efforts outlined here serve to enhance the user experience for all audiences, furthering our vision of a more equitable and accessible world.

9.2. Fairness in SWIPE results

The data presented in this section highlights that the results of SWIPE do not show significant differences or strong effect sizes based on gender and age variables. It is important to note that AssessFirst only requests personal information necessary for the appropriate use of the platform. For instance, we do not collect information about religious, political, or sexual orientation. Regarding age, we only ask for date of birth to ensure it does not impact how questions are handled. Moreover, the variables analysed below do not play any role in the calculation of results within the AssessFirst solution. Our commitment to protecting user privacy and promoting inclusivity is reflected in our data practices.

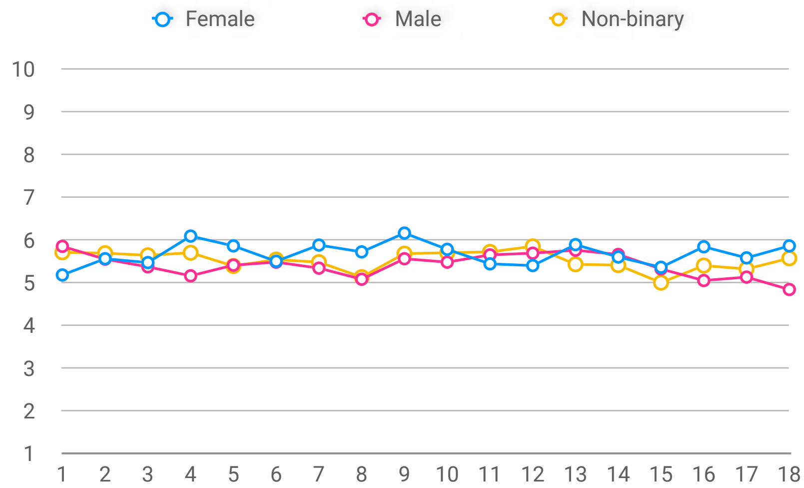
9.2.1. Fairness regarding gender

Historically, data on personality facets have shown minimal gender differences, suggesting that males and females exhibit similar behaviours. Despite popular belief, the differences that do exist are often exaggerated, and most psychological and cognitive attributes between genders are largely comparable. For instance, Janet Shibley Hyde's pioneering meta-analysis in 2005 hypothesised gender similarity, demonstrating that differences between genders are negligible or very weak in 78% of cases, particularly in psychological factors. Similarly, Ethan Zell, Zlatan Krizan, and Sabrina Teeter's study in 2015 reinforced these findings, with 84% overlap in the distribution of scores between genders and weak or very weak effects in 85% of cases. Recent studies suggest that some gender differences may arise from how data is organised, with gender differences becoming more apparent when several indicators that differ according to gender are combined to produce global typicality scales (Eagly & Revelle, 2022). Moreover, recent research has demonstrated that algorithms based on personality-related data have no adverse impact, with an average impact ratio of .91 (Kubiak, Baron, & Niesner, 2023; Efremova, Kubiak, Baron, & Frasca, in press).

However, it is important to note that these results should not overshadow the fact that there may be slight, natural differences between men and women in certain specific personality facets. However, if such differences exist, they remain almost negligible or small. Females tend to have slightly higher scores than males on traits such as agreeableness and neuroticism, whilst men tend to have slightly higher scores on extroversion and conscientiousness traits. However, these differences in scores between genders are often small, and there is also a great deal of individual variability (Schmitt et al., 2008; Weisberg et al., 2011; Costa et al., 2001; Lippa, 2010; Kajonius & Johnson, 2018). Moreover, as indicated by effect size indices from various studies, these differences are generally modest and tend to only concern a few facets. For instance, the most pronounced effects are found on:

- Compassion (d = .45), Politeness (d = .36), Emotional volatility (d = .30) and Withdrawal (d = .40), and on traits Agreeableness (d = .48) and Neuroticism (d = .39) (Weisberg, Deyoung & Hirsh, 2011) ;
- Anxiety (d = .56), Altruism (d = .51), Modesty (d = .45) and Sympathy (d = .57), and on traits Agreeableness (d = .58) and Neuroticism (d = .40) (Kajoniusa & Johnson, 2018) ;
- Anxiety (d = .43), Assertiveness (d = .27) and Altruism (d = .32) (Costa, Terracciano & McCrae, 2001) ;

To summarise, previous research has shown that certain personality facets may be more sensitive to gender differences, specifically Assertiveness, Anxiety, and Compassion. Given these findings, it is likely that SWIPE may also show similar effects with similar effect sizes. To confirm this, AssessFirst conducted gender equity studies in April 2023 (N=3,001), including 1,624 females, 985 males, and 392 non-binary individuals. The results are presented in Graph 9.1 and Table 9.1 and are based on Cohen's d-effect size. A value of d ≈ .0 indicates no effect, a value of d ≈ .3 corresponds to a weak effect, d ≈ .5 corresponds to a medium effect, and d ≈ .8 corresponds to a strong effect.



Graph 9.1. Average scores for the 18 facets according to gender.

Overall, the average scores in all facets of SWIPE are close to the theoretical average of 5.5, ranging from 5 to 6. However, as previously mentioned, there are slight differences in scores for certain facets, particularly in Compassion (.93), Anxiety (.78), Assertiveness (.67), and Emotional Volatility (1.01). Whilst these differences may be partly theoretically explained, they are likely also influenced by the composition of the sample used in the analysis.

Facets	Cohen's d	Effect size
Assertiveness	-.35	Weak
Energy level	.01	-
Sociability	.05	-
Compassion	.44	Weak
Respectfulness	.21	Very weak
Trust	.01	-
Greed avoidance	.25	Very weak
Modesty	.34	Weak
Sincerity	.26	Very weak
Aesthetic sensitivity	.14	-
Creative imagination	-.10	-
Intellectual curiosity	-.14	-
Organisation	.06	-
Productiveness	-.03	-
Responsibility	.02	-
Anxiety	.40	Weak
Depression	.21	Very weak
Emotional volatility	.54	Medium

Table 9.1. Cohen's d and effect size according to gender (Male/Female).

The effect sizes observed in SWIPE are consistent with the existing literature, which suggests that there are some differences between males and females in personality facets related to agreeableness and neuroticism. However, it is important to note that these differences are rare and mostly weak or very small. Furthermore, these differences may be partly due to the sampling effect, as individuals who choose to participate in online research may have specific personality tendencies (Valentino, Zhirkov, Hillygus & Guay, 2020; Marcus & Schütz, 2005). To conclude, the results of SWIPE suggest that there are no major differences between males and females on the 18 facets measured. Therefore, SWIPE can be considered gender-equitable.

To further examine gender equity, we present Tables 9.2 and 9.3, which compare the effect sizes between females and non-binary individuals, and males and non-binary individuals, respectively. It is worth noting that the effects observed in both tables are rare and mostly very weak.

Facets	Cohen's d	Effect size
Assertiveness	-.28	Very weak
Energy level	-.05	-
Sociability	-.08	-
Compassion	.18	-
Respectfulness	.22	Very weak
Trust	-.02	-
Greed avoidance	.19	-
Modesty	.32	Weak
Sincerity	.21	Very weak
Aesthetic sensitivity	.04	-
Creative imagination	-.14	-
Intellectual curiosity	-.22	Very weak
Organisation	.19	-
Productiveness	.09	-
Responsibility	.19	-
Anxiety	.23	Very weak
Depression	.12	-
Emotional volatility	.16	-

Table 9.2. Cohen's d and effect size according to gender (Female/Non-binary).

Facets	Cohen's d	Effect size
Assertiveness	-.07	-
Energy level	-.07	-
Sociability	.13	-
Compassion	-.25	Very weak
Respectfulness	-.00	-
Trust	.03	-
Greed avoidance	.06	-
Modesty	.03	-
Sincerity	.05	-
Aesthetic sensitivity	.10	-
Creative imagination	.03	-
Intellectual curiosity	.08	-
Organisation	-.13	-
Productiveness	-.12	-
Responsibility	-.17	-
Anxiety	.17	-
Depression	.08	-
Emotional volatility	.37	Weak

Table 9.3. Cohen's d and effect size according to gender (Male/Non-binary).

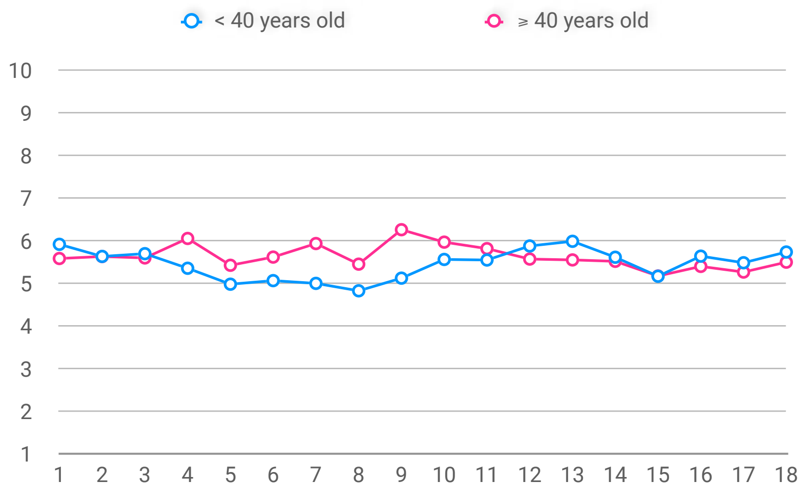
9.2.2. Fairness regarding age

The answer to whether personality changes with age is not straightforward. Decades of research in psychology suggest that personality is relatively stable over time, but it is not entirely immutable. A meta-analysis of longitudinal studies conducted by Bleidorn et al. in 2022 supports this claim:

- Young adulthood is the most critical life stage for personality development (Arnett, 2000; Roberts & Mroczek, 2008; Roberts & DelVecchio, 2000; Roberts & Davis, 2016). It is in early adulthood that facets crystallise, and most facets undergo pronounced changes. Specifically, throughout childhood and adolescence, facets are relatively unstable, but during the transition to young adulthood, they become increasingly stable, with the achievement of a peak of stability around age 25 years old;
- Estimates of personality stability peak around age 25, level off in mid-adulthood, and remain stable or possibly decline slightly in old age (see also Roberts, Walton & Viechtbauer, 2006; Soto et al., 2011).

In summary, the literature suggests that personality becomes highly stable in young adulthood. Whilst some changes can occur in adulthood, they typically involve increases in agreeableness (Roberts, Walton, & Viechtbauer, 2006) and emotional stability. Therefore, mean scores for personality facets are not expected to differ significantly by age, indicating fairness. However, there may be a slight tendency for older age groups to have higher mean scores on facets related to agreeableness and lower mean scores on facets related to emotional stability (Roberts, Walton, & Viechtbauer, 2006).

SWIPE's latest age equity studies, based on Cohen's d effect size analysis, were conducted in April 2023 (N=306), on a sample with an average age of 40 years (σ = 10.97). On the basis of this average age, two comparison groups were chosen: people whose age is < 40 years (N = 155) and people whose age is ⩾ to 40 years (N = 151). The results are presented in Graph 9.2 and Table 9.4. A value of d ≈ .0 indicates no effect, a value of d ≈ .3 corresponds to a weak effect, d ≈ .5 corresponds to a medium effect, and d ≈ .8 corresponds to a strong one.



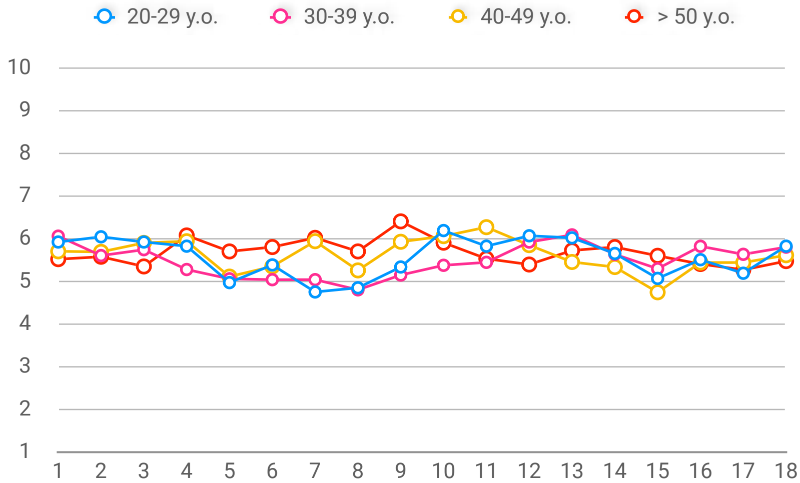
Graph 9.2. Mean scores for the 18 facets according to age (2 classes).

Overall, the average scores for all facets fall between 5 and 6, which is in line with the theoretical average of 5.5. The largest differences are observed in facets related to Agreeableness and Humility traits: Compassion (.69), Sincerity (1.13), Greed avoidance (.93), and Modesty (.62). However, it is important to note that these differences may be influenced by the small sample size.

Facets	Cohen's d	Effet size
Assertiveness	.16	-
Energy level	.00	-
Sociability	.05	-
Compassion	-.30	Weak
Respectfulness	-.19	-
Trust	-.28	Very weak
Greed avoidance	-.46	Weak
Modesty	-.32	Weak
Sincerity	-.49	Weak
Aesthetic sensitivity	-.19	-
Creative imagination	-.11	-
Intellectual curiosity	.14	-
Organisation	.18	-
Productiveness	.04	-
Responsibility	-.00	-
Anxiety	.11	-
Depression	.09	-
Emotional volatility	.12	-

Table 9.4. Cohen's d and effect size according to age (2 classes).

The effect sizes observed in the study confirm the existing literature on the subject, indicating that there are no significant age-related differences. Whilst the effect sizes are weak or very weak, some tendencies related to agreeableness facets (higher average scores for those ⩾ 40 years old) and emotional stability facets (lower average scores for those < 40 years old) are consistent with the initial hypotheses proposed in previous research (Roberts, Walton & Viechtbauer, 2006). It should be noted, however, that (1) these effects are rare and only apply to 5 out of the 18 facets measured by SWIPE, (2) the effect sizes are mainly small, and (3) they may be due to sampling effects since the study only involved a sample of N = 306 individuals. Overall, the results suggest that there are no meaningful differences between individuals ⩾ 40 years old and those < 40 years old on the 18 facets measured by SWIPE, and thus the results are considered fair across different age groups.



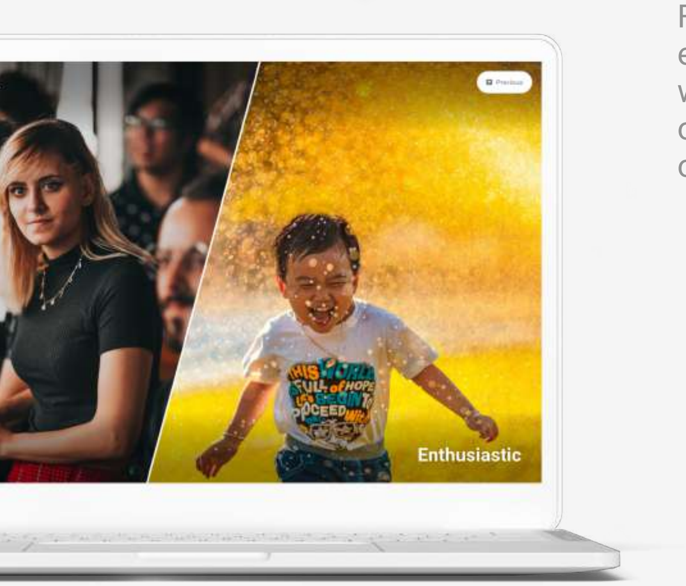
Graph 9.3. Mean scores for the 18 facets according to age (4 classes).

To conduct a more thorough analysis of age equity, four age categories were considered: (1) 20 to 29 years old, (2) 30 to 39 years old, (3) 40 to 49 years old, and (4) 50 years old and over. The average scores for all categories were found to be between 5 and 6, which is close to the theoretical average of 5.5. It is also worth noting that there were no significant differences between the age categories that were large enough to be considered strong effects.

9.2.3. Conclusion

The results of the AssessFirst SWIPE assessment do not show any significant differences based on gender or age categories. This suggests that the assessment is fair and does not discriminate against any particular gender or age. The effect sizes observed in gender-related differences were very weak or weak and were only observed in a few facets. It is also possible that these effects could be explained by other factors or sampling biases, and they can be conceptually justified. Overall, the results suggest that the AssessFirst SWIPE assessment is a reliable and fair tool for assessing personality facets across different genders and age groups.

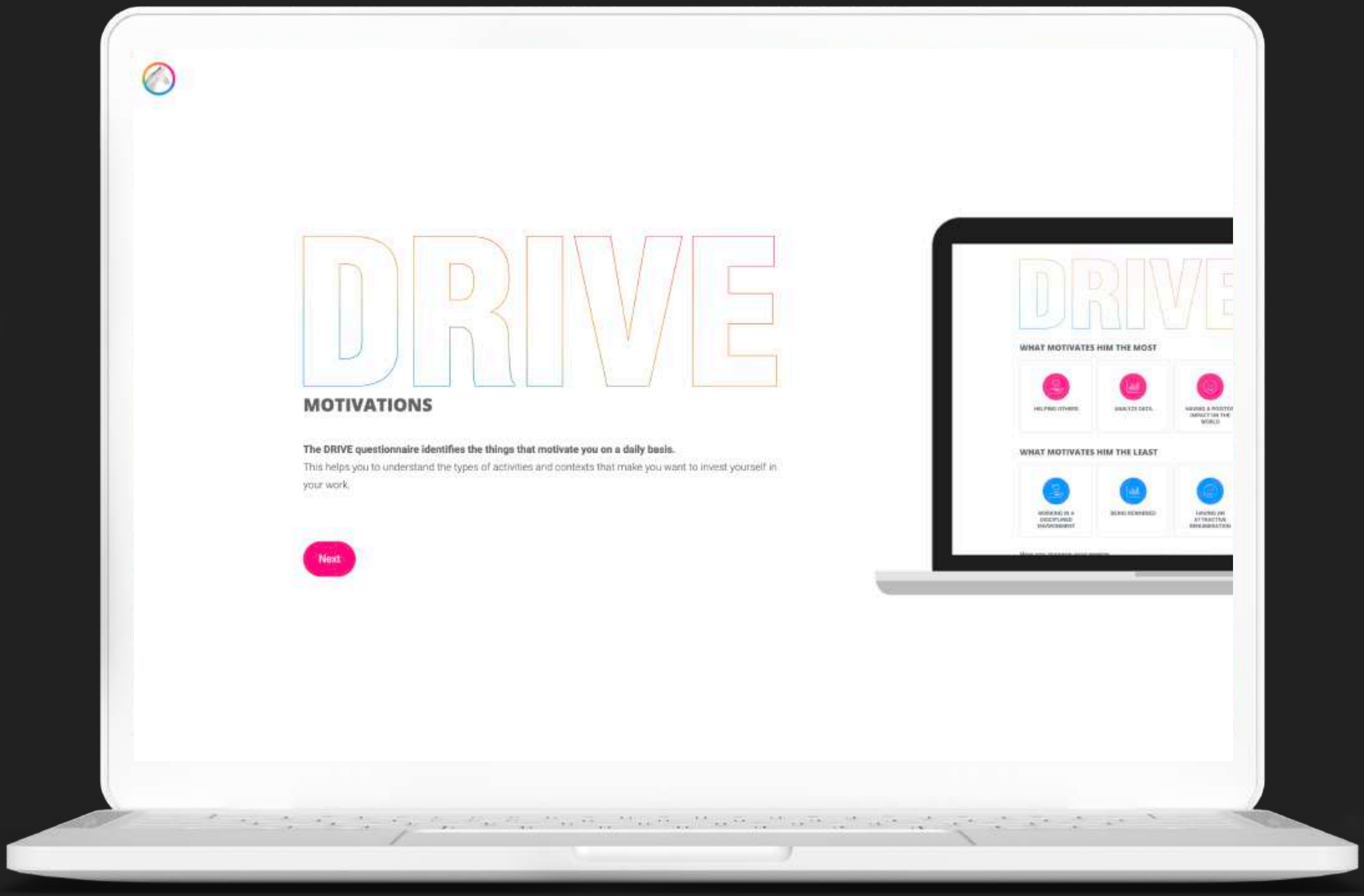
Summary of fairness



Fairness in the context of a personality assessment refers to the extent to which the assessment is designed and administered in a way that is fair and unbiased for all individuals who take it, regardless of their demographic characteristics such as gender, sexual orientation, race, ethnicity, or culture.

.08 The mean Cohen's d for the gender variable indicates that there is no significant effect on most personality facets.

.08 The mean Cohen's d for the age variable indicates that there is no significant effect on most personality facets.



Identify peoples' motivational factors.

DRIVE is a brief assessment designed to assess motivational factors, providing insights into what drives a person's satisfaction and engagement. It takes only 10 minutes to complete and helps to predict how well individuals will adapt to a particular role and work environment.

DRIVE

1. Introduction

DRIVE is a concise motivational assessment that helps to identify the factors influencing individual satisfaction and engagement. It can be completed in 10 minutes and is used to predict an individual's integration ability into a specific context and role. The assessment consists of 90 items that measure 20 dimensions of motivation. DRIVE was developed in 2014 by the Science team at AssessFirst.

2. Development history

The assessment of motivations, which emerged as a predominant field of interest in psychology in the last century, asks the fundamental question: what are the forces that drive a person to dedicate themselves and engage in an activity? This question has captivated pioneers in the field who sought to map the complex landscape of human motivations. Among them, Abraham Maslow laid the groundwork with his famous hierarchy of needs in 1954, closely followed by David McClelland with his theory of needs in 1961, and Frederick Herzberg with his two-factor theory in 1968. These thinkers set milestones by attempting to formulate universal models of motivation, influencing contemporary understanding of what drives us to act in the professional world.

Recent studies have revealed that motivation is a far more complex concept than what initial unidimensional theories suggested. It cannot be attributed to a single factor, but rather results from a set of interdependent factors. In 2005, Latham and Pinder provided a comprehensive analysis of the existing literature, leading to significant findings:

- Job satisfaction is a significant predictor of performance (Judge et al., 2001).
- Workplace motivation reduces the willingness to quit (Williams et al., 2001).
- Workplace motivation reduces absenteeism and enhances engagement (Wegge et al., 2007).

These conclusions mark a turning point in the understanding of motivational dynamics in the workplace and emphasise the need for human resource management policies that recognise the multidimensionality of motivation.

AssessFirst's ambition with the development of DRIVE was to build upon academic discoveries regarding motivation and create a practical tool for the benefit of organisations and individuals. The idea was to enable everyone to make informed decisions about their careers. DRIVE is thus presented as an analytical instrument aimed at identifying the key factors that influence employee satisfaction and engagement at work. In doing so, AssessFirst sought to provide companies with a lever to optimise the well-being and productivity of their teams by personalising the approach based on each individual's motivations.

The development of DRIVE by AssessFirst's Science team began in 2014, marking the start of an intensive phase of research and innovation. This period was characterised by a thorough review of scientific literature on motivation, the evaluation of existing measurement instruments, and a detailed study of the elements that define engagement and professional success. In early 2016, DRIVE was launched, offering the result of these efforts: a tool designed to precisely discern what propels individuals to success in their roles. This represents significant work that reflects AssessFirst's commitment to providing evidence-based solutions to businesses and professionals to promote success and well-being in the professional environment. DRIVE complemented AssessFirst's suite of assessment tools, building on a decade and a half of expertise. This new tool leveraged proven methodologies and best practices developed by the company, including diagnostic accuracy, user-friendliness, and data and results interpretation. Furthermore, with the integration of item response theory, an advanced psychometric model, DRIVE offers advanced settings for a precise and personalised assessment of motivation and workplace engagement factors. This signifies a shift to a more sophisticated approach tailored to the contemporary talent management needs of businesses.

3. Theoretical foundations

Over the past fifty years, psychology has significantly contributed to our understanding of motivation. At AssessFirst, we have had the privilege of drawing upon a rich body of theoretical reflections and empirical studies to craft our DRIVE assessment. Among the wealth of research, some have stood out for their relevance and alignment with our ultimate goal: deciphering the essential elements that cultivate employee satisfaction and engagement. These works, which have directly influenced our approach, have enabled us to build a tool that not only measures but also enlightens and guides toward better professional alignment. Here is a summary of the research that has been pivotal in this process.

3.1. Self-determination theory

The design of DRIVE drew strong inspiration from Deci and Ryan's Self-Determination Theory, which holds a central place in the field of modern psychology. This theory, widely recognised for its academic authority, posits that an individual's sustainable motivation cannot be solely ensured by extrinsic rewards. For a person to remain engaged in an activity in the long term, they must tap into their intrinsic motivations. By integrating these principles, DRIVE seeks to identify and assess these intrinsic motivations, providing a more nuanced and profound insight into what truly motivates employees at work, beyond simple material incentives. This enables DRIVE users to understand and cultivate a work environment that promotes self-sustaining motivation, essential for engagement.

According to the Self-Determination Theory, intrinsic motivation is fueled by the satisfaction of three fundamental psychological needs. Firstly, the need for competence, which refers to an individual's ability to succeed and master the tasks assigned to them. Secondly, the need for autonomy, which emphasises the importance of freedom of choice and the individual's capacity to overcome obstacles by showing initiative and creativity. Finally, the need for relatedness, which highlights the desire to be in relationships with others, to feel a sense of belonging to a group, and to be supported by peers. DRIVE incorporates these needs into its analysis to provide a framework for individuals and organisations to understand and enhance intrinsic motivations.

DRIVE reflects the Self-Determination Theory in its structure. Out of the twenty dimensions assessed by the test, eighteen correspond to intrinsic motivations, reflecting the belief that these factors are predominant for long-term engagement. However, recognising the importance of not excluding the role of extrinsic motivations, DRIVE also includes two extrinsic dimensions. This decision is based on the meta-analysis by Cerasoli, Nickin & Ford in 2014, which demonstrated that in certain professional circumstances, these extrinsic factors can have a positive influence. DRIVE is designed to assess the satisfaction of the three fundamental needs at the core of the Self-Determination Theory, ensuring that the tool is grounded in a deep understanding of what drives employees in their workplace.

While the Self-Determination Theory provides a robust framework for understanding and assessing motivation, it is important to note that it is not specifically designed to identify all the specific factors that promote individual well-being beyond the professional context. In other words, this theory focuses on auditing the level of motivation related to autonomy, competence, and relatedness but does not claim to explore all the conditions of personal flourishing that can be affected by various and diverse factors not covered by this model.

3.2. Motives, Values, Preferences Inventory

The development of DRIVE has been greatly influenced by the work of Joyce and Robert Hogan, who have made a significant impact in the field of psychology. Their rigorous analysis spans eight decades of research, resulting in a model that has been recognised for excellence for nearly twenty years. This model stands out for its ability to cross-integrate various existing theories and experimental studies while being particularly well-suited for professional contexts, as it was specifically designed for this domain. That's why the ten dimensions of the Motives, Values, Preferences Inventory (MVPI) are indirectly integrated into the foundations of DRIVE, enriching the tool with a broad spectrum of motivational factors.

3.3. Multidimensional theory of person-environment fit

The meticulous exploration of recent advancements in psychology has allowed our team to push the boundaries of traditional understanding of motivation. This investigation has led us to a more sophisticated conceptualisation, integrating the modeling of various levels of alignment or "fit" between an individual and their professional environment. This multidimensional concept of "fit" is central in DRIVE because it goes beyond assessing motivation itself; it also seeks to determine how the alignment between personal aspirations, skills, values, and the work environment can optimise workplace engagement and satisfaction.

The majority of research on workplace satisfaction and engagement focuses on the impact of the connection between an individual and their work environment. The significant research conducted by Edwards and Billsberry, published in 2010, highlighted the importance of various levels of "fit" in a professional context. These levels of "fit," inspired by the model developed by Jansen and Kristof-Brown in 2006, include:

- ‘*Person-vocation fit*’: this refers to the alignment between an individual's interests, values, skills, and abilities, and the requirements or rewards of a profession or career. A good fit in this context means that the person is well-suited to the tasks, values, and expectations of the profession they have chosen.

- ‘*Person-organisation fit*’: this concept focuses on the compatibility between an individual's values and the culture, norms, and values of an organistion. A good "fit" means that the person shares similar values to those of the organisation, which can lead to increased job satisfaction and a higher likelihood of success within the company.

- ‘*Person-group fit*’: this refers to the match between an individual and a specific work group within the organisation, including teams or departments. A good fit can lead to improved collaboration, communication, and teamwork.

- ‘*Person-job fit*’: this term describes the relationship between an individual's skills, experiences, and expectations, and the requirements of the job they hold. A good fit can result in better performance, greater job satisfaction, and reduced turnover.

- '*Person-person fit*': This refers to the interpersonal relationship and compatibility between individuals at work, whether with colleagues, superiors, or subordinates. A good alignment can improve communication, mutual respect, and support within the organisation.

This nuanced understanding is incorporated into DRIVE: allowing for the measurement and analysis of these various forms of alignment to better understand and promote the engagement and satisfaction of individuals in their environment and work. The five levels of "fit" identified in research significantly influence three key indicators in the professional context: employee engagement, their intention to stay or leave their position, and their overall job satisfaction. With the stated goal of strengthening these indicators, DRIVE was developed based on these proven models. This strategy has guided our approach to create a tool capable of addressing and improving the alignment between employees and their work environment in a targeted manner, with the ultimate goal of fostering a professional environment where engagement and satisfaction are optimised.

3.4. DRIVE's dimensions

Dimensions	Description
Create new things	Need to express creativity in their work
Excel everyday	Need to be confronted with ambitious challenges
Worry about aesthetics	Need to take care of the presentation of their work
Analyse data	Need to work on factual elements and conduct analyses
Meet new people	Need to have a people-centered job
Have clearly defined tasks	Need to make progress on short-term work
Worry about quality	Need to perform reliable and precise work
Having influence	Need to have decision-making power
Having autonomy	Need for freedom of action
Working as part of a team	Need to collaborate with others in their work
Having a positive impact on the world	Need to work for a company that has a positive impact on the world
Working in a fun environment	Need for social stimulation and a relaxed work environment
Working in a reassuring environment	Need for a stable environment and long-term employment
Working in a disciplined environment	Need to adhere to established rules and principles
Maintaining personal balance	Need to have personal time, not just for work
Being rewarded	Motivated by the prospect of a reward
Having an attractive remuneration	Need to earn a significant income and accumulate value
Seeks competition	Need for victories, success, and achieving goals
Helping others	Need to support others
Being recognised by others	Need to be appreciated and valued by others

Table 3.1. Dimensions of motivations measured by DRIVE.

4. Development of DRIVE

The design of DRIVE is centered around four fundamental pillars:

- **Scientific foundation:** Each factor assessed by DRIVE is based on robust scientific principles, supported by conclusive research publications.
- **Professional relevance**: The measured factors resonate directly with the realities of today's working world and are designed to be immediately applicable in professional contexts.
- **Accessibility of the assessment**: The assessment process is crafted to be intuitive and seamless, ensuring the assessment can be completed with ease.
- **Clarity of results**: Data derived from DRIVE is presented in a way that is immediately understandable and actionable for users.

These principles have been critical in developing a tool that precisely meets the requirements and expectations of businesses while being universally applicable, regardless of the user's profile.

4.1. Choice of the dimensions

The development of DRIVE was enriched by a thorough exploration of available knowledge, encompassing scientific publications, business models, existing assessment tools, and studies on professional success. Building on the previously defined core principles, the team meticulously selected 25 motivational factors that met these stringent criteria. These factors were chosen for their robustness in both scientific literature and empirical practice, thereby ensuring that DRIVE is grounded on a foundation of evidence-based data and practical applications in the professional world.

Upon reviewing existing tools for measuring motivation, we noted that most analyse a limited spectrum, often fewer than ten criteria. While this may be sufficient to establish a foundation for understanding motivation, it falls short in detecting the nuances of individual motivational needs. Analogous to the Big Five personality traits, which provide a robust framework for understanding an individual's general characteristics, these broad domains require more detailed analysis to gain a deep understanding of personal attributes. DRIVE is thus designed to go beyond this surface-level assessment, offering a detailed and nuanced exploration of motivational factors. On the other hand, our analysis identified extremely detailed tools that evaluate more than 30 criteria. These instruments provide a comprehensive understanding of motivational drivers, but their complexity makes the results difficult and time-consuming to utilise. This is in direct conflict with one of our fundamental principles: to provide clear and easily actionable information. Therefore, DRIVE was intentionally created to balance the depth of analysis with ease of use, ensuring that users gain relevant insights without being overwhelmed by excessive complexity.

The selection of the initial 25 factors for DRIVE represents a careful balance between oversimplification and unnecessary complexity. We have meticulously crafted clear and distinct definitions for each of these factors to ensure they embody specific and independent concepts. This approach has resulted in an assessment framework that is both content-rich and user-friendly, thus avoiding the pitfalls of a tool that is either too elementary or overly burdensome to interpret.

4.2. Items' conception

On the solid foundation formed by the factors and their definitions, a team of five occupational psychologists independently crafted the items of DRIVE. Their writing was guided by three key directives to ensure the tool's accessibility and relevance:

- **Semantic clarity**: Each item had to be written in plain and clear language to be understandable by anyone, regardless of their familiarity with psychological terms.
- **Semantic precision**: It was imperative to use unambiguous terms to avoid any ambiguity that could compromise the reliability of the responses.
- **Contemporary relevance**: The items needed to reflect current and real situations in the work environment to ensure that the tool is grounded in modern professional reality.

These guidelines have led to the creation of items that not only precisely measure motivations but are also intuitive and directly applicable for end-users, thus enhancing the practical value of DRIVE. Additionally, in writing the items for DRIVE, a dual strategy was adopted, inspired by the findings of Brown & Maydeu-Olivares in 2012. For each measured concept, items were formulated to capture the essence of the factor (positive items) as well as items designed to represent the inverse or antithesis of the same concept (negative items). This binary approach aims to enrich the dimensionality of the assessment and minimise response biases. By offering this alternation between positive and negative statements, DRIVE increases the accuracy and reliability of the assessment, allowing for a more balanced measurement of motivations.

For DRIVE, an exhaustive item development process was established, with the creation of approximately 750 items across the 25 identified factors, which equates to roughly 30 items per factor. To validate their quality and relevance, these items were meticulously reviewed by an external group of 20 observers not involved in the creation phase. These observers were tasked with linking each item to a specific factor, thus ensuring a clear and unequivocal association between the item and the concept it is supposed to measure. Furthermore, to ensure clarity of formulation, each item was rated on a scale from 1 (not clear) to 5 (very clear). This rigorous process aims to eliminate any ambiguity and ensure that each item contributes significantly to the precise measurement of the corresponding motivational factor. The final selection of items for the first version of the DRIVE assessment was based on a combined criterion of relevance and clarity. Only items that were consistently associated with their respective factor by at least 80% of observers and which scored an average rating above 4 out of 5 in terms of clarity were included. This meticulous methodology ensures that each item included in DRIVE is both precisely related to the concept it is intended to measure and formulated in a way that is easily understandable, thus ensuring the fidelity and effectiveness of the assessment from its launch.

4.3. Development

To validate the selected items, an initial normative version of the DRIVE assessment was developed and made available online. Respondents were allowed to rate each item on a scale from 1 (not at all motivating) to 5 (very motivating). To prevent participant fatigue, which could introduce bias into the results, the assessment was split into two separate versions, each containing 280 items. This division aimed to lighten the load for respondents and ensure the integrity of the data collected, as taking a 560-item assessment at once could impact the quality of responses due to fatigue.

Each 280-item version of the DRIVE assessment was administered to a sample of about 300 voluntary individuals. These participants, representing the active European population aged between 25 and 55 years and working full-time, provided the initial data for a preliminary analysis. The responses gathered allowed for a structural validity study, with the results documented in the section dedicated to the assessment's validity. This study aimed to confirm the organisation and coherence of the motivation factors as measured by DRIVE, ensuring that the tool adheres to psychometric standards and meets the requirements of scientific validity.

Following the initial statistical analysis, it was noted that certain dimensions of the DRIVE assessment did not offer sufficient statistical reliability for two main reasons:

- The lack of internal consistency for some factors meant that even with a substantial number of items intended to describe the concept, it wasn't possible to identify a consistent set of items that would be strongly correlated with each other, which is necessary for reliably assessing a unique concept.
- Conceptual redundancy was observed when certain factors were too similar conceptually or were opposites, making it difficult to distinguish them as independent entities. Consequently, decisions were made to eliminate or merge some factors by consolidating their most relevant items and removing the least discriminating ones.

These adjustments are crucial to ensure that each dimension of DRIVE assesses a distinct and relevant aspect of motivation, thereby increasing the validity and usefulness of the assessment for practical applications in the professional environment.

After revision, the DRIVE assessment was refined to 20 factors with a total of 372 carefully selected items. To improve accessibility and simplicity, a forced-choice technique was employed, pairing items based on their social desirability. This method requires respondents to choose between two propositions, reducing social desirability bias and simplifying the response process. This approach is based on experience and best practices in psychometrics and was chosen because it proves to be most effective in facilitating administration while preserving the quality of the collected data. The final version of DRIVE, structured around the "forced-choice" format, consisted of 186 pairs of items. This version was administered online to a panel of 347 individuals, all professionals aged 25 to 55 years, working full-time. The analysis of the responses allowed for an assessment of the balance between the items in each pair. The goal was to achieve a distribution of responses close to parity, ideally with 50% of respondents selecting each of the two propositions. A distribution gap was, however, acceptable, with a tolerance set at 60% for one proposition and 40% for the other. This process ensures that each pair of items is balanced, thereby avoiding a potential bias towards a more socially desirable option and contributing to the precision of DRIVE's measurements.

For items that did not meet the target distribution of 50/50 or the acceptable range of 60/40, a revision was undertaken to adjust their level of attractiveness, with the aim of correcting the observed asymmetry. If the revision threatened to compromise the validity of the original item, it was then discarded to preserve the fidelity of the instrument. This refinement process resulted in a revised version of the DRIVE assessment, consisting of 166 pairs of items. This new version was administered to a group of 341 individuals, selected to reflect the same demographic characteristics as the previous panels. This methodical adjustment aimed to perfect the measurement and ensure that the assessment remained both precise and accessible to respondents.

In the final version of the DRIVE assessment, only the 90 pairs of propositions with the best psychometric qualities were retained. This stringent selection was aimed at focusing the assessment on the most effective and reliable items while ensuring a reduced completion time. This strategy fulfills the commitment to offer a tool that is both robust and respectful of the respondents' time, thereby facilitating its integration into professional contexts where time is a crucial factor. The result is an optimised assessment that efficiently captures motivational factors without imposing an excessive burden on participants.

4.4. Format

The adoption of a two-option forced-choice assessment format for DRIVE is based on significant advantages this type of approach offers:

- Reduction of social desirability bias: The forced-choice structure is known for its effectiveness in minimising social desirability bias. In a selection context where respondents might be tempted to present themselves in the best light, controlling this bias is crucial (Christiansen et al., 2005; Jackson et al., 2000; Martin et al., 2001; Vasilopoulos et al., 2006).
- Ease and speed of administration: Compared to a normative version, the forced-choice format allows for time savings. Observations have shown that a forced-choice assessment can be completed 30% more quickly, making it more convenient for both the organisation and the candidate.
- Elimination of response biases: Normative formats can often lead to central tendency biases (the tendency to choose neutral options) and extremity biases (a preference for extreme options). The forced-choice format neutralises these tendencies by forcing respondents to make a more considered choice between two alternatives.

These reasons support the decision to opt for a forced-choice assessment, thus maximising the validity and efficiency of DRIVE in professional assessment processes. The use of the term "forced-choice" rather than "ipsative format" to describe the DRIVE assessment is related to psychometric specifics and the purpose of the obtained scores. The ipsative format, characterised by a constant sum of scores, indeed poses several psychometric issues:

- Score relativity: in an ipsative format, the obtained scores reflect a personal norm, indicating the priority a person gives to certain characteristics over others. They do not represent an absolute measure but a relative comparison among different preferences or tendencies (Closs, 1996; Hicks, 1970; C. E. Johnson et al., 1988).
- Construct validity: the interdependencies among items in an ipsative format can limit construct validity. Covariance between items can lead to statistical distortions, making it difficult to distinguish between constructs (Baron, 1996; Cornwell & Dunlap, 1994).
- Internal reliability: the reliability, or internal consistency, of an ipsative test is often questioned. Ipsative scores are not independent of each other, complicating the comparison of internal reliability between ipsative and normative formats (Meade, 2004).

Conversely, the forced-choice format used by DRIVE aims to minimise these issues by not imposing a constant sum of scores, thus allowing for a more nuanced evaluation of individual preferences without the psychometric constraints associated with the ipsative format. This offers advantages in terms of the precision and validity of measurements, which is crucial for professional selection applications where the interpretation of scores must be as faithful as possible to the reality of the traits being evaluated. Furthermore, the decision to apply Item Response Theory (IRT) for calculating DRIVE's results rather than Classical Test Theory (CTT) was a strategic choice to overcome the intrinsic limitations of forced-choice assessment formats. IRT provides a precise, individualised measure by considering the specific difficulty of each item and the respondent's ability. This results in trait estimations that are not affected by sample or test item variations. Additionally, research by Brown & Maydeu-Olivares in 2012 supports the idea that IRT can indeed transcend the biases traditionally associated with the forced-choice format, thereby offering a fairer comparison capacity and more reliable data interpretations. This advanced approach ensures that DRIVE provides not only sophisticated and robust evaluations of motivations but also tailored ones.

5. Validity

How can we determine if an assessment accurately measures what it claims to measure? How can we ensure that each scale is measured correctly and that the results of the assessment have the intended meaning? These questions are answered through validation studies. The purpose of validating an assessment is to confirm that it measures the intended construct and to determine the accuracy of the results obtained from it. In the past, validity was defined as the correlation between a score on an assessment and an external criterion that measured either the same construct or a construct that was supposed to be related to the construct associated with the score. To establish and ensure the validity of an assessment, several types of validity must be examined. The validity studies of DRIVE cover the following types of validity:

- **Content validity**: refers to the extent to which the items of an assessment semantically represent an adequate sample of the content domain being measured. This means that the items should be directly related to the construct they are intended to measure and also cover all the main aspects of that construct;
- **Construct validity**: refers to the degree to which the assessment accurately measures the psychological construct or dimension it is designed to assess. This type of validity is established through various analyses, such as item-dimension saturation, inter-dimension correlation, and distribution parameters;
- **Predictive validity**: the predictive validity of a personality assessment measures its ability to predict a target variable, such as job performance or turnover. In other words, it assesses whether the results of the personality assessment can be used to predict future outcomes in the workplace.

5.1. Content validity

Content validity is a critical criterion in the development of psychometric assessments. It measures the relevance of the content and its ability to cover all dimensions of a theoretical concept. It assesses not only the representativeness of the items with respect to the underlying theoretical model but also the quality of their formulation. Unlike methods that rely on statistical analyses, content validity requires a rational approach to establish a connection between the construct to be measured and the questions asked. A common method to assess this validity involves soliciting the opinions of experts in the field targeted by the inventory. These experts judge the relevance of the items and may also be invited to associate each item with the theoretical dimensions they represent, without preconceptions. This approach, which was favored by the Science team, provides a more informative validation process. This method was specifically chosen during the development of DRIVE because it best matched the standards and requirements of the time, thus ensuring a solid and reliable foundation for the tool.

The content validity study of DRIVE was conducted with ten judges. They were selected to participate in this study. Each judge received a first document (document A) detailing the 180 items of the DRIVE inventory and a second document (document B) presenting the definitions of the 20 dimensions assessed by this inventory. The instruction presented to the judges was as follows:

> *The DRIVE inventory comprises 20 dimensions. Each dimension has 9 items. For each of the items presented in document A, you must select from document B the dimension to which it relates. An item can only belong to a single dimension.*

The following table shows, for each dimension, the number of items (out of the 9 items for that dimension) for which inter-judge agreement is very high (between 90 and 100%), high (between 80 and 90%), correct (between 70 and 80%), and insufficient (between 0 and 70%). The study conducted on the content validity of DRIVE reveals very positive results. Inter-judge agreement—which measures the consensus among experts on the relevance of items to the dimensions they are supposed to measure—is extremely encouraging. With a very high rate of inter-judge agreement for the majority of the dimensions (7.25 out of the dimensions evaluated), and a high level for a small proportion (1.05 dimension), this indicates a strong consistency in the experts' evaluation. Moreover, the level of concordance is considered correct for less than one dimension (0.7), with no items judged as having an insufficient degree of agreement. These data demonstrate that the items of DRIVE are almost unanimously perceived as relevant and representative of the dimensions they are intended to measure. Therefore, it can be confidently stated that the content validity of DRIVE is robust, reflecting a rigorous construction and a reliable methodology that enhance its value as a psychometric assessment tool.

Dimensions	91-100%	81-90%	71-80%	0-70%	Total
Create new things	9	0	0	0	9
Excel everyday	8	1	0	0	9
Worry about aesthetics	7	1	1	0	9
Analyse data	7	1	1	0	9
Meet new people	7	1	1	0	9
Have clearly defined tasks	7	2	0	0	9
Worry about quality	6	2	1	0	9
Having influence	8	0	1	0	9
Having autonomy	9	0	0	0	9
Working as part of a team	8	1	0	0	9
Having a positive impact on the world	7	2	0	0	9
Working in a fun environment	7	0	2	0	9
Working in a reassuring environment	7	1	1	0	9
Working in a disciplined environment	8	0	1	0	9
Maintaining personal balance	8	1	0	0	9
Being rewarded	6	2	1	0	9
Having an attractive remuneration	5	2	2	0	9
Seeks competition	8	1	0	0	9
Helping others	6	2	1	0	9
Being recognised by others	7	1	1	0	9
Mean	**7.25**	**1.05**	**.70**	**0**	**9**

Table 5.1. Inter-judge agreement.

5.2. Construct validity

Construct validity refers to whether an assessment instrument measures the intended theoretical construct and not something else. It is closely related to other aspects of validity, as any evidence of validity contributes to understanding the construct validity of a test. The importance of construct validity lies in the fact that it influences the interpretation of test scores. If a test claims to measure a specific dimension, it is crucial to ensure that it actually measures that dimension. Otherwise, any interpretation of the scores would be incorrect and could lead to biased decisions. However, construct validity is not limited to simply looking at whether the assessment is measuring a specific dimension. It involves a comprehensive investigation to determine whether the interpretations of the test results are consistent with the theoretical and observational terms that define the construct (Cronbach & Meehl, 1955).

There is no single method for determining construct validity, but rather different methods and approaches must be combined. In order to assess the construct validity of DRIVE, we have utilised three complementary methods: item-dimension saturation and inter-dimension correlation (Thurstone, 1947; Bollen, 1989; McDonald, 2013), and the presentation of distribution parameters (Fisher, 1921, 1922).

- **Item-dimension saturation** refers to the correlation between an item and the total score of the dimension or factor to which it belongs. In other words, if an item is designed to measure a particular dimension, it should be closely associated with other items that measure that dimension. Thus, the higher the correlation between an item and the dimension, the more strongly the item is related to that dimension and therefore more valid. For item-dimension saturation, a value of .40 or higher is generally considered satisfactory and adequate. This suggests that the item measures the dimension it is supposed to measure (Campbell & Fiske, 1959; Nunnally, 1978; Hair, Black, Babin, & Anderson, 2010). Saturation below .40 may be acceptable if supported by theoretical justification;

- **Inter-dimension correlation** assesses the relationship between scores of different factors or dimensions measured by a test. If two dimensions are expected to be distinct and independent, then they should have weakly correlated scores. On the other hand, if the dimensions are closely related or overlapping, the scores should be more strongly correlated. There is no universal threshold for inter-dimension correlation. It is generally desirable for the dimensions to be independent, although there may be some moderate correlations between the dimensions that are justified by the underlying theoretical model.

- **The distribution parameters** correspond to the statistical characteristics of the distribution of scores in the assessments. These parameters include measures such as the mean, standard deviation, skewness, and kurtosis, which provide information about the shape, centre, and variability of the distribution. They make it possible to identify atypical scores, explore individual differences in the distribution of scores, and to better interpret the results. For example, high levels of skewness or kurtosis may indicate non-normal distributions, which could affect the interpretation of the results and the use of certain statistical tests. Therefore, it is important to examine distribution parameters in addition to other aspects of validity when assessing the quality of an assessment.

5.2.1. Item-dimension saturation

The most recent item-dimension saturation studies for DRIVE were conducted with a sample size of N = 4450, on the primary items measuring each dimension, indicate favorable results. The saturation, or the correlation of items with their intended dimension, ranged from .52 to .83, and the average saturation per dimension ranged from .63 to .69. These results confirm that the item-dimension saturations for DRIVE are satisfactory and adequate, suggesting that the items are well-correlated with the dimensions they are intended to measure, which is a positive indication of the inventory's construct validity.

Dimensions	i1	i2	i3	i4	i5	i6	i7	i8	i9	Mean
Create new things	.69	.59	.65	.71	.71	.65	.78	.62	.60	**.67**
Excel everyday	.63	.64	.73	.60	.76	.66	.56	.61	.68	**.65**
Worry about aesthetics	.60	.63	.58	.66	.65	.60	.647	.72	.62	**.63**
Analyse data	.57	.83	.76	.67	.71	.65	.63	.63	.65	**.68**
Meet new people	.70	.69	.7	.62	.65	.60	.64	.64	.78	**.67**
Have clearly defined tasks	.68	.67	.61	.76	.69	.75	.71	.61	.61	**.68**
Worry about quality	.65	.64	.64	.73	.61	.59	.60	.71	.59	**.64**
Having influence	.71	.65	.64	.72	.61	.66	.73	.62	.57	**.66**
Having autonomy	.61	.69	.61	.67	.66	.60	.68	.70	.67	**.65**
Working as part of a team	.67	.7	.75	.68	.71	.68	.69	.65	.66	**.69**
Having a positive impact on the world	.63	.66	.61	.65	.67	.52	.77	.63	.62	**.64**
Working in a fun environment	.65	.64	.71	.61	.66	.67	.74	.68	.67	**.67**
Working in a reassuring environment	.62	.57	.74	.62	.62	.61	.74	.60	.70	**.65**
Working in a disciplined environment	.60	.65	.60	.63	.67	.59	.67	.61	.68	**.63**
Maintaining personal balance	.65	.78	.62	.60	.67	.63	.61	.67	.75	**.66**
Being rewarded	.66	.67	.67	.75	.64	.65	.59	.69	.65	**.66**
Having an attractive remuneration	.69	.7	.61	.7	.68	.62	.64	.61	.71	**.66**

Table 5.2. Item-dimension saturation.

5.2.2. Inter-dimension correlation

The most recent correlation studies of DRIVE's inter-dimension aspects were conducted with a sample size of N = 4,450. The dimensions show weak correlations, thus supporting an acceptable level of consistency. The most correlated pairs are "Having an attractive remuneration" and "Having a positive impact on the world" (r = -.45), "Having an attractive remuneration" and "Being rewarded" (r = .43), and "Having clearly defined tasks" and "Having influence" (r = -.44).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		.21	-.10	.04	.07	-.25	-.30	.28	.06	.08	.12	-.12	-.24	-.10	-.33	-.11	-.26	-.01	-.10	-.25
2			.05	.10	.02	-.04	.08	.12	-.06	-.15	-.02	-.32	-.03	-.26	-.10	-.15	-.09	.20	-.21	-.33
3				.05	.05	.23	.31	-.19	-.07	.04	.07	-.22	.03	-.21	.21	-.37	-.12	-.15	.05	-.28
4					-.26	.15	.18	.08	-.19	-.14	-.04	-.30	.07	-.09	.08	-.09	-.04	.08	-.15	-.07
5						-.13	-.23	-.19	-.17	.37	.22	.06	-.22	-.03	-.06	-.16	-.29	-.32	.26	-.06
6							.36	-.44	-.07	-.08	-.09	-.09	.26	.11	.35	-.22	-.01	-.08	-.13	-.15
7								-.28	-.09	-.27	-.25	-.24	.23	-.22	.32	-.12	.08	.15	-.10	-.08
8									.05	-.11	-.13	-.05	-.16	-.14	-.36	.13	.01	.21	-.21	.02
9										-.13	-.04	.17	-.15	.13	-.13	-.04	.05	-.12	-.03	-.12
10											.23	.05	-.21	-.04	-.11	-.23	-.30	-.26	.32	-.02

11												.11	-.25	.01	-.03	-.29	-.45	-.40	.28	-.04
12													-.12	.28	-.15	.03	-.13	-.33	.12	.25
13														.07	.26	.10	.26	.12	-.24	-.07
14															.07	-.05	-.01	-.24	.03	.04
15																-.10	.05	-.03	.02	-.22
16																	.43	.29	-.20	.28
17																		.36	-.28	.06
18																			-.33	-.01
19																				.05
20																				

Table 5.3. Inter-dimension correlation for DRIVE.

5.2.3. Distribution parameters

The distribution of scores obtained from a motivation assessment is an essential aspect of the assessment's construct validity. The way scores are distributed for each motivation dimension can provide vital information about how the test measures that dimension and how the scores are interpreted. Our analysis of distribution parameters focuses on five primary parameters that are necessary:

- the **mean,** which is an indicator of the central tendency of the scores in the distribution;
- the **median**, which is a measure of central tendency that represents the value that divides a distribution in half, with 50% of the scores above and 50% below the median. Unlike the mean, the median is less sensitive to extreme scores and is a more robust measure of central tendency;
- the **standard deviation**, which is a measure of the dispersion of scores around the mean. It is calculated by taking the square root of the variance of the scores;
- **skewness**, which is a measure of the symmetry of a distribution. It is calculated by comparing the frequency of scores to the left and right of the mean. If the distribution is perfectly symmetric, the skewness is zero. If the distribution is skewed to the left, the skewness is negative. If the distribution is skewed to the right, the skewness is positive;
- **kurtosis,** which is a measure of the degree of peakedness or flatness of a distribution compared to a normal distribution. A normal distribution has a kurtosis of 0. If a distribution is more peaked than a normal distribution, its kurtosis value is positive, and if it is less peaked, its kurtosis value is negative.

Expectations for distribution parameters depend on the context and the measurement instrument used. However, in general, here is what is expected for "good" distribution parameters:

- the mean should be close to the median value: this indicates that the distribution is symmetric. If the mean is significantly different from the median, this may indicate an asymmetry in the distribution;
- the standard deviation should be reasonable and large enough to capture individual differences in the measured dimension, but not so large as to dilute the differences between individuals. In general, one would expect the standard deviation to be around 2 for the 10-point personality scale;
- the asymmetry (skewness) should be close to 0 (symmetrical distribution). If the skewness is significantly different from 0, this may indicate an asymmetry in the distribution;
- the kurtosis should be close to 0 (normal distribution). If the kurtosis is significantly different from 0, the distribution is either flatter or more peaked.

The latest studies on the distribution parameters of DRIVE were conducted in 2023 (N= 48,449).

Dimension	Mean	Median	SD	Skewness	Kurtosis
Create new things	5.71	6.00	2.02	0.05	-0.25
Excel everyday	5.45	5.00	1.93	0.17	0.21
Worry about aesthetics	5.33	5.00	1.98	-0.10	0.05
Analyse data	5.45	5.00	2.03	0.04	-0.66
Meet new people	5.65	6.00	2.03	-0.04	-0.59
Have clearly defined tasks	5.27	5.00	1.84	0.02	-0.38
Worry about quality	5.10	5.00	1.91	0.12	-0.15
Having influence	5.69	6.00	1.90	-0.04	-0.32
Having autonomy	5.52	6.00	1.86	-0.15	-0.08
Working as part of a team	5.71	6.00	1.92	0.09	0.14
Having a positive impact on the world	5.93	6.00	2.09	0.00	-0.43
Working in a fun environment	5.92	6.00	1.90	-0.05	-0.33
Working in a reassuring environment	5.12	5.00	2.01	-0.01	-0.43
Working in a disciplined environment	5.52	6.00	1.96	-0.10	-0.24
Maintaining personal balance	5.12	5.00	1.89	-0.05	-0.22
Being rewarded	5.29	5.00	2.04	-0.15	-0.27
Having an attractive remuneration	5.17	5.00	2.06	0.10	-0.70
Seeks competition	5.33	5.00	1.82	0.10	-0.13
Helping others	5.60	6.00	1.90	0.12	0.26
Being recognised by others	5.69	6.00	1.83	0.07	-0.24
	5.48	*5.50*	*1.95*	*.01*	*-.24*

Table 5.4. Distribution parameters for each dimension.

The distribution parameters are consistent with the expected standards. The normality of distributions can be interpreted through indicators such as the overlap of means and medians, and with the aid of skewness and kurtosis coefficients that are close to 0.

5.3. Predictive validity

Predictive validity of a motivation assessment is the measure of its ability to predict a target variable, such as performance or turnover. In other words, it is about whether the assessment results can be used to forecast future performance. Evidence of predictive validity is particularly relevant when one wishes to infer, from a assessment score, an individual's position on another criterion variable that is independently assessed at a later date. Studies of predictive validity related to DRIVE are presented in a dedicated guide.

5.4. Conclusion

The validation of a motivation assessment is essential to ensure that the obtained measurements are accurate. In this study, we have examined the content validity, construct validity, and predictive validity of DRIVE. Our analyses indicate that:

- (1) DRIVE encompasses the theoretical constructs it is supposed to measure.
- (2) The assessment is well-structured and demonstrates good measurement homogeneity.
- (3) The dimensions are predictive of job success.

Overall, the results meet the most demanding psychometric standards and demonstrate the validity of DRIVE. However, to further explore the psychometric qualities of DRIVE, its reliability must be considered. In this sense, a assessment must be both valid and reliable to be used in professional decision-making (recruitment, mobility, etc.). Indeed, a valid but unreliable assessment would mean that the test accurately measures what it is supposed to, but the individual scores are inconsistent. On the other hand, a valid and reliable assessment consistently measures what it is intended to measure: in other words, it consistently hits the bullseye. Evidence of reliability is presented in the following chapter.

6. Reliability

How can you determine if the results of an assessment are reliable? How can you ensure that the assessment produces consistent results when asking the same questions to the same person at different times? The answers to these questions can be obtained through the study of reliability. Whilst validity provides information on an assessment's ability to measure what it intends to measure, reliability measures whether the measurement is consistent and reliable every time the same assessment is completed by the same person. In short, the reliability of an assessment measures its consistency or stability over time and aims to determine if an assessment produces similar results when asking the same questions to the same person at different times or to similar people. Therefore, the objective of reliability is to ensure that the obtained results are dependable and accurate. The reliability of a assessment can be evaluated in two different and complementary ways:

- **Internal consistency,** which is a statistical measure used to assess the reliability of a psychometric test. It evaluates the homogeneity or similarity of different test items that are intended to measure the same psychological dimension. In other words, internal consistency assesses whether multiple items that are designed to measure the same thing produce similar scores;

- **Test-retest reliability**, which is a method used to assess the reliability of a measurement by measuring the same variable at two different points in time. This approach enables the assessment of the temporal stability of the measurement and the estimation of the proportion of total variance attributable to measurement error. Test-retest reliability is frequently utilised in longitudinal studies or to evaluate the stability of a test over a specific period.

6.1. Internal consistency

The latest Cronbach's alpha studies for DRIVE were conducted with a sample size of 332. With the exception of four dimensions that have α values ranging from .62 to .69, all alpha coefficients are above .70, indicating adequate reliability results for the DRIVE assessment. Moreover, these results should be contextualised considering the forced-choice structure of DRIVE, which renders Cronbach's alpha a less than optimal indicator in this scenario.

Dimension	α	Dimension	α
Create new things	.77	Having a positive impact on the world	.72
Excel everyday	.71	Working in a fun environment	.72
Worry about aesthetics	.68	Working in a reassuring environment	.72
Analyse data	.70	Working in a disciplined environment	.77
Meet new people	.75	Maintaining personal balance	.81
Have clearly defined tasks	.87	Being rewarded	.71
Worry about quality	.77	Having an attractive remuneration	.75
Having influence	.74	Seeks competition	.82
Having autonomy	.69	Helping others	.62
Working as part of a team	.79	Being recognised by others	.67

Table 6.1. Alpha coefficients for DRIVE's dimensions.

6.2. Test-retest reliability

Test-retest reliability is a measure of the temporal consistency of a assessment or measurement scale. It involves administering the same assessment to a group of participants at two different points in time, with a time interval between the two administrations. The correlation between the two sets of results is then calculated to determine the reliability of the test. A high correlation indicates that the participants' scores are stable over time, suggesting that the assessment can be considered reliable.

The latest test-retest reliability studies for DRIVE were conducted with a sample size of 232 individuals. The table below provides the test-retest reliability of DRIVE administered 3 months apart. The 20 dimensions of the DRIVE assessment demonstrate good test-retest reliability with coefficients ranging from .67 to .83 and a median reliability coefficient of .74. This stability of scores over a significant period highlights DRIVE's ability to provide consistent and reliable measures of motivations, which is a critical asset for assessments in a professional context where long-term decisions are based on these data.

Dimension	Pearson r	Dimension	Pearson r
Create new things	.76	Having a positive impact on the world	.71
Excel everyday	.72	Working in a fun environment	.77
Worry about aesthetics	.69	Working in a reassuring environment	.70
Analyse data	.73	Working in a disciplined environment	.69
Meet new people	.75	Maintaining personal balance	.83
Have clearly defined tasks	.79	Being rewarded	.74
Worry about quality	.74	Having an attractive remuneration	.71
Having influence	.77	Seeks competition	.75
Having autonomy	.67	Helping others	.77
Working as part of a team	.74	Being recognised by others	.71

Table 6.2. Test-retest reliability for DRIVE's dimensions.

7. Sensitivity

Sensitivity, also called discrimination, refers to the ability of an assessment to distinguish between people with a high level on a dimension and people with a low level. It, therefore, reflects the ability of the assessment to identify the uniqueness of each individual. A sensitive motivation assessment can identify the subtle differences between people, and help to understand their behaviours with more precision and discrimination. Sensitivity thus refers to the ability of the assessment to correctly identify people who possess the characteristic being measured and to avoid false positives (people who do not possess the characteristic, but who are identified as such by the assessment).

The first attempts to measure discrimination were based on cumulative scales (Guttman, 1944; Walker, 1931; Loevinger, 1948; Loevinger, 1953, cited by Hankins, 2008). However, Ferguson was one of the first to propose conceptualising discrimination in the form of a coefficient. In this sense, if there is a maximum number of possible differences in a sample, the discrimination coefficient corresponds to the ratio between the number of differences actually observed and this maximum number of differences. This coefficient called the delta δ of Ferguson (Ferguson, 1949; Kline, 2000), is thus the ratio between the differences observed between people and the number of maximum possible differences. It is intended to be a direct and non-parametric index of the degree of distinction made by an instrument between individuals. If no difference is observed, then $\delta = 0$. If all possible discriminations are observed, then $\delta = 1$. Generally, a normal distribution should have excellent discrimination, where $\delta \geqslant .9$ (Ferguson, 1949). Weaker discriminations are expected for leptokurtic distributions (because these distributions fail to discriminate around the mean) and skewed distributions (because these fail to discriminate at one end of the distribution). Demonstrating excellent discrimination between the scales of an assessment requires a $\delta \geqslant .9$ for each scale. For example, in a recent adaptation study of the BFI-2 in Russian, Kalugin, Shchebetenko, Mishkevich, Soto, and John (2021) showed that all scales had strong discriminations.

The latest sensitivity studies of DRIVE, based on the calculation of Ferguson's delta (δ), were conducted in 2023 with a sample size of 12,386. All δ coefficients are above .9, indicating excellent discrimination of the measurement scales. In other words, this means that the assessment is capable of accurately detecting individual differences in the motivations of the tested individuals, and that it is sensitive to individual variations in the measured dimensions. The results are presented in the table below.

Dimensions	δ	Dimensions	δ
Create new things	.97	Having a positive impact on the world	.98
Excel everyday	.97	Working in a fun environment	.96
Worry about aesthetics	.97	Working in a reassuring environment	.96
Analyse data	.97	Working in a disciplined environment	.98
Meet new people	.97	Maintaining personal balance	.97
Have clearly defined tasks	.96	Being rewarded	.99
Worry about quality	.97	Having an attractive remuneration	.98
Having influence	.96	Seeks competition	.94
Having autonomy	.97	Helping others	.98
Working as part of a team	.98	Being recognised by others	.96

Table 7.1. Ferguson's delta for DRIVE's dimensions.

8. Fairness

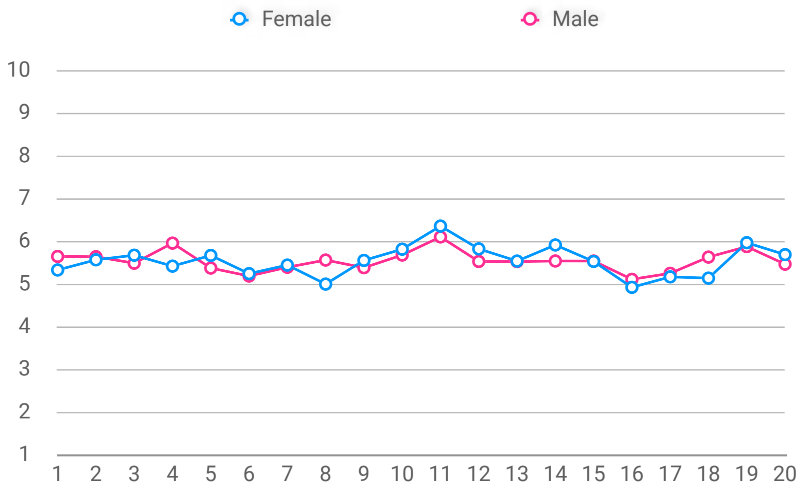
Fairness in the context of a motivation assessment refers to the extent to which it is designed to be fair and unbiased for all individuals, regardless of their origin, gender, sexual orientation, race, or culture. In other words, a fair assessment should be objective and impartial towards all individuals who take it, without any bias or discrimination against any particular group. Our teams take every measure to ensure the fairness of our assessments and predictive analyses, and we ensure that the use of our algorithms in decision-making processes does not lead to discrimination through any unforeseen algorithmic biases. Additionally, in the development of our assessments, equity studies focus on two areas: (1) ensuring the accessibility of the assessment, and (2) ensuring equity in the results of the assessment.

8.1. Fairness in DRIVE's results

The data presented in this section highlights that the results of DRIVE do not show significant differences or strong effect sizes based on gender and age variables. It is important to note that AssessFirst only requests personal information necessary for the appropriate use of the platform. For instance, we do not collect information about religious, political, or sexual orientation. Regarding age, we only ask for date of birth to ensure it does not impact how questions are handled. Moreover, the variables analysed below do not play any role in the calculation of results within the AssessFirst solution. Our commitment to protecting user privacy and promoting inclusivity is reflected in our data practices.

8.1.1. Fairness regarding gender

The latest gender equity analyses for DRIVE were conducted in 2022 on a sample (N = 332,587) consisting of 51% men and 49% women.



Graph 8.1. Average scores for the 20 dimensions according to gender.

Overall, all means are between 5 and 6, which is close to the theoretical average of 5.5. Therefore, there are no major differences between the results of men and women across the 20 dimensions measured by DRIVE. The results are thus equitable across genders. These conclusions are further supported by the effect sizes, presented below.

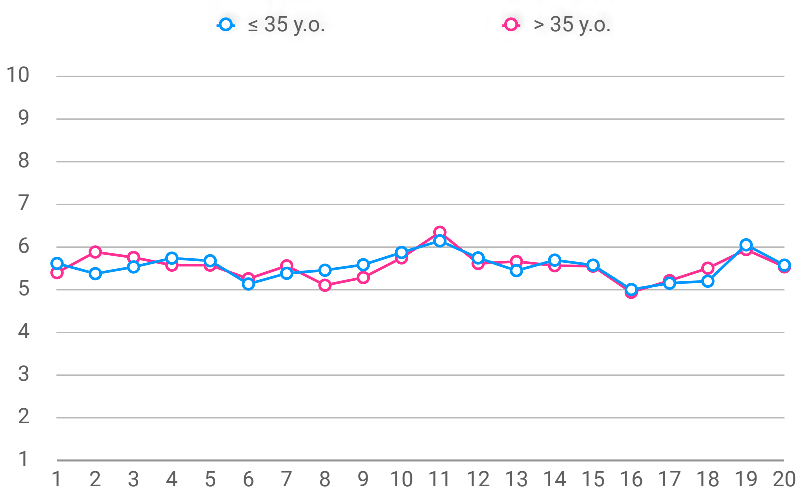
The effect sizes highlighted here support the previous conclusion. However, it is important to note that: (1) these effects are rare, (2) they are exclusively very weak effects, (3) they are potentially inherent to a sampling effect. In conclusion, although some effects are highlighted, there are no significant differences between the results of men and women on the 20 DRIVE dimensions. Therefore, DRIVE's results are equitable across genders.

Dimensions	*d*	Effect	Dimensions	*d*	Effect
Create new things	.16	-	Having a positive impact on the world	-.12	-
Excel everyday	.04	-	Working in a fun environment	-.15	-
Worry about aesthetics	-.09	-	Working in a reassuring environment	-.01	-
Analyse data	.26	Very weak	Working in a disciplined environment	-.18	-
Meet new people	-.15	-	Maintaining personal balance	.00	-
Have clearly defined tasks	-.03	-	Being rewarded	.08	-
Worry about quality	-.03	-	Having an attractive remuneration	.04	-
Having influence	.29	Very weak	Seeks competition	.27	Very weak
Having autonomy	-.09	-	Helping others	-.04	-
Working as part of a team	-.06	-	Being recognised by others	-.12	-

Table 8.1. Cohen's d for each dimension.

8.1.2. Fairness regarding age

The latest age equity analyses for DRIVE were conducted in 2022 on a sample (N = 64,675) composed of 42% of individuals under 35 years of age, and 58% of individuals over 35 years of age. To compare results, the sample was divided into two age groups from the age of 35 (due to the average age of the sample being 33.3 years, with a standard deviation of 10.2 years).



Graph 8.2. Average scores for the 20 dimensions according to age.

Overall, all means are between 5 and 6, which is close to the theoretical average of 5.5. Consequently, there are no major differences between the results of individuals based on their age across the 20 dimensions measured by DRIVE. The results are therefore equitable according to age. These conclusions are further substantiated by the effect sizes, which are presented in the subsequent section.

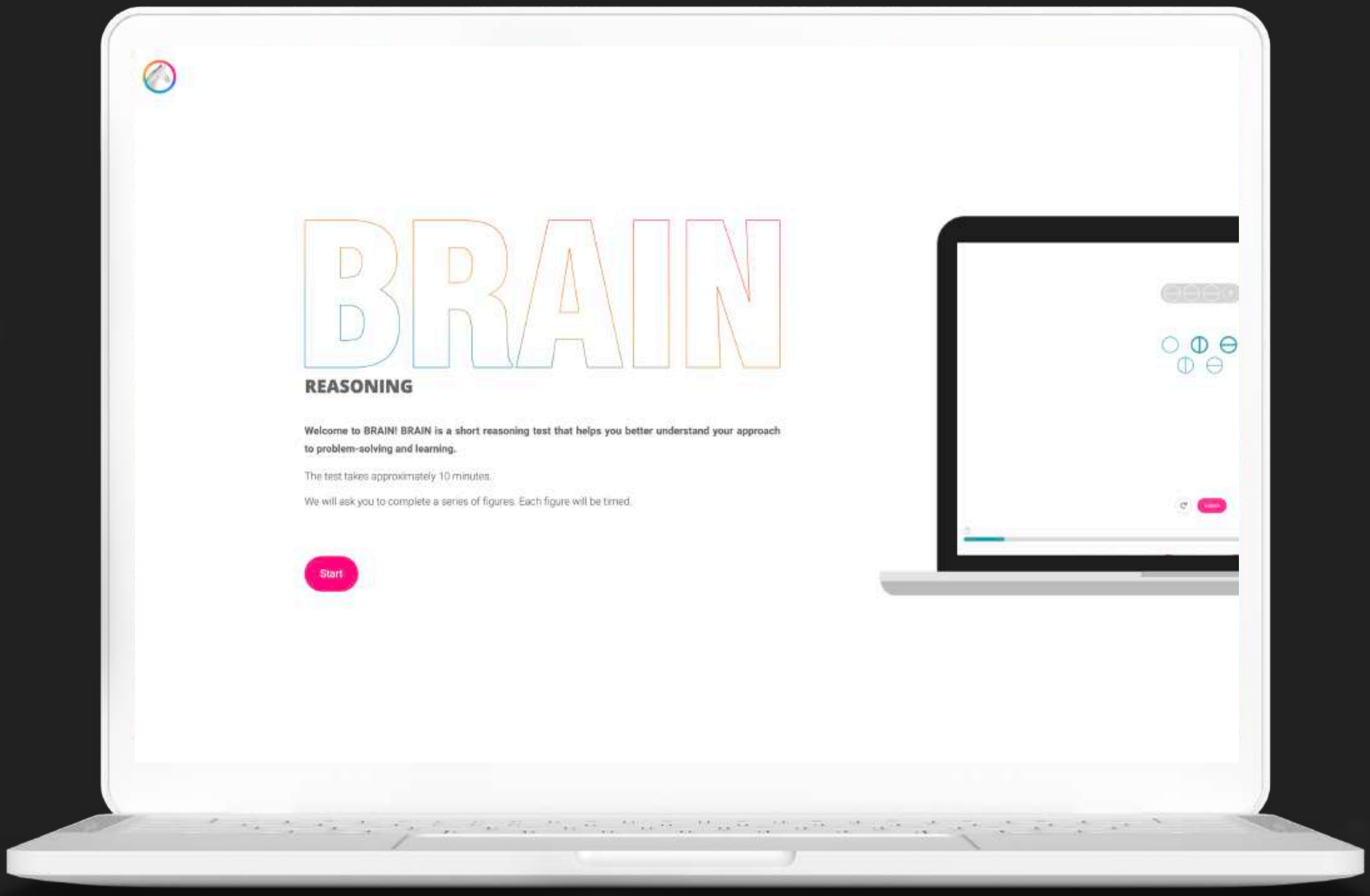
The effect sizes highlighted here reinforce the previous conclusion. It is important to note that: (1) these effects are infrequent, (2) they are exclusively very weak effects, (3) they may be potentially inherent to a sampling effect. In conclusion, while some effects are noted, there are no significant differences between the results of younger and older individuals on the 20 DRIVE dimensions. Thus, the results of DRIVE are equitable across age groups.

Dimensions	*d*	Effect	Dimensions	*d*	Effect
Create new things	.11	-	Having a positive impact on the world	.09	-
Excel everyday	.25	Very weak	Working in a fun environment	.07	-
Worry about aesthetics	.11	-	Working in a reassuring environment	.11	-
Analyse data	.08	-	Working in a disciplined environment	.06	-
Meet new people	.05	-	Maintaining personal balance	.01	-
Have clearly defined tasks	.06	-	Being rewarded	.03	-
Worry about quality	.09	-	Having an attractive remuneration	.03	-
Having influence	.18	-	Seeks competition	.16	-
Having autonomy	.15	-	Helping others	.05	-
Working as part of a team	.06	-	Being recognised by others	.02	-

Table 8.2. Cohen's d for each dimension.

8.1.3.Conclusion

Whether in terms of gender or age categories, there are no major differences between the results obtained by different groups on the AssessFirst DRIVE assessment. In summary, the obtained results support the hypothesis that the proposed assessment does not discriminate against a particular gender or age. The most pronounced effects pertain to gender and only a few dimensions. Also, these mentioned effects are very weak.



Dive into problem-solving

BRAIN is a brief reasoning assessment designed to gauge an individual's general reasoning ability and decision-making process. Adaptive and gamified, it can be completed in just 10 minutes, striking a balance between the accuracy of the assessment and the quality of the user experience.

BRAIN

1. Introduction

BRAIN is a concise reasoning assessment that assesses a person's general reasoning capabilities and decision-making process. Adaptive, gamified, and designed to be completed in 10 minutes, BRAIN harmonises the reliability of measurement with the quality of user experience. The assessment consists of 8 to 12 items that evaluate various dimensions including the overall potential for reasoning, decision speed, preferred task types, learning style, ability to handle complexity, decision accuracy, risk factors, and behavioral style. BRAIN was developed in 2020 by the Science team at AssessFirst.

2. Development history

Technological advancements have rendered work transient and intricate, with most jobs now demanding skills to solve non-routine and dynamic problems. As such, an individual's reasoning abilities are paramount. Intelligence is defined as a general capacity for understanding and learning quickly. Linda Gottfredson, a professor of psychology at the University of Delaware, describes it as the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience. However, other theories suggest viewing intelligence through specific and multiple competencies, arguing that an individual might excel in one area (such as numerical manipulation) but not as much in others: Psychologist Joy Paul Guilford, for instance, identifies up to 150 different types of intelligences. Despite the debate, research has long shown that while there are specific forms of reasoning, they are strongly interrelated and tied to a general factor of intelligence, known as the *g* factor. In the workplace, the *g* factor is one of the most explanatory variables of an employee's performance, especially in roles involving high levels of complexity. Its significance appears to be less in jobs where decisions and problems are simpler and more routine. Analysis of 20,000 studies involving 5 million individuals by Nathan Kuncel, Deniz Ones, and Paul Sackett at the University of Minnesota demonstrates that cognitive abilities can predict 25% of a future employee's performance, notably because they enable quicker acquisition of the knowledge necessary for the task (Kuncel, Ones, and Sackett, 2010). Additionally, Steffanie Wilk, Laura Desmarais, and Paul Sackett, researchers in psychology and management, conclude that individuals with strong cognitive abilities are more likely to move up the hierarchy (Wilk, Desmarais, and Sackett, 1995).

Despite these findings, HR decision-makers often still prioritise other criteria, such as experience, in the recruitment of candidates or in managing career progression. However, John Hunter, a psychology professor at Michigan State University, has shown that an individual's level of reasoning is, on average, three times more predictive of future performance than experience. A more recent study even suggests that experience has no significant correlation with success in the role for new employees (Van Iddekinge, Frieder, and Roth, 2019). In a rapidly changing work environment, the majority of individuals need to learn and adopt new methods of production. Establishing a recruitment policy that accounts for candidates' reasoning abilities is therefore a prerequisite for optimising employee performance, their adaptation to the future, and economic growth. While this approach may seem elitist, it does not exclude those with lower cognitive abilities from employment. Indeed, although automation will inevitably restructure the job market in the coming years, particularly affecting unskilled professions, historical structural changes show that for every job eliminated by technology, new activities have been created. For instance, a study by McKinsey & Company concluded that for every job lost due to the evolution of the internet, 2.4 jobs have been created. Furthermore, simply recruiting intelligent candidates does not guarantee that they will be effective and committed; it is the match between the job requirements and the candidate's abilities that improves efficiency and limits turnover. Steffanie Wilk and Paul Sackett explain that if an employee's level of reasoning exceeds the requirements of the position, they may be more likely to leave (Wilk and Sackett, 2006).

Using a reasoning assessment tool allows for recruitment based on a criterion that truly explains success, but also to precisely know the level of complexity a candidate can handle, and what kind of tasks to assign them: simple, familiar tasks for those with less intellectual ease, or new and strategic issues for those capable of higher-level reasoning. To meet recruiters' needs and integrate smoothly into recruitment processes, reasoning tests have evolved significantly, particularly through the application of game-based assessment theories. This new format of cognitive tests allows for a reliable measure of candidates' reasoning levels much faster, enhances user engagement and motivation, and can even reduce their anxiety during the test (Burgers, Eden, Van Engelenburg, and Buningh, 2015). Moreover, to remain relevant, companies must revise their recruitment criteria. In this regard, a candidate's level of reasoning is a highly predictive data point of their future success and must be integrated into recruitment strategies, alongside analyses of personality and motivation. This paradigm shift also calls for an evolution in education systems, emphasising problem-solving and critical analysis. Such efforts are necessary to enable everyone to express their talents and find their place in the society of tomorrow.

Launched in 2020 by AssessFirst, BRAIN is a cutting-edge tool that addresses the need for measuring general intelligence or the *g* factor, a key indicator of professional performance across a variety of roles. This tool stands out by aligning with current trends and market demands through three core principles:

- Optimal accessibility, ensuring smooth use on mobile devices, catering to the mobility of modern users.
- An adaptive approach that adjusts the difficulty of questions in real-time based on the candidate's responses, providing a personalised and precise assessment of their abilities.
- A design aimed at engaging and motivating candidates, making the assessment process not only more enjoyable but also more immersive.

These innovations position BRAIN at the forefront of psychometric advancements, in line with the evolving needs of clients and candidates in today's professional landscape. Thus, BRAIN enhances the testing conditions, but its goal remains the same: to acquire a precise, decontextualised, and cross-functional indicator of an individual's reasoning capability.

The BRAIN test is a general evaluation designed without segmentation into various sections. At the heart of this unique experience, candidates are invited to solve a series of logical items within a set time. These items are not resolved through traditional multiple-choice questions, but rather by constructing responses using interactive elements within a virtual environment. This creative format allows for a more dynamic assessment of cognitive abilities. Moreover, BRAIN relies on adaptive formats (Computerised Adaptive Testing or CAT), which continuously adjust the difficulty level of items according to the candidate's performance: a first item of intermediate difficulty is followed by items that vary in difficulty based on previous responses, increasing if the answer is correct and decreasing if not. The test concludes when the candidate stabilises their responses at a consistent difficulty level, allowing for a precise measurement of general intelligence tailored to each individual.

The BRAIN assessment offers numerous advantages:

- Adaptive evaluation: the adaptive models (CAT) used in BRAIN create a custom test for each candidate, as the items presented are tailored to their level and actual performance on previous items. This allows for (1) quicker determination of a candidate's level, (2) an improved user experience as the candidate does not feel like they are failing, (3) reduced possibility of answer leakage and cheating since no candidate has the same items. Adaptive models thus reduce the time required for the test by offering items suited to the candidate's real level and performance. On average, a candidate will encounter between 8 and 12 items in the new version of BRAIN, corresponding to about 10 minutes (compared to 38 items and 21 minutes in the older version).
- Equitable by design: BRAIN is developed with a mobile-first approach, meeting the growing demand of candidates to apply and complete recruitment tests on mobile devices. Additionally, the neutral material used in BRAIN (i.e., material free of verbal elements and based on colors unrelated to the ability to solve items) makes the test accessible to people with disabilities.
- Intercultural: the neutral material used by BRAIN minimises the need for language adaptations.
- Gamified: while not a game, BRAIN employs similar interaction and motivation mechanisms: real-time feedback, answer construction, adapted item levels, and personalised instructions. This test format enhances candidate engagement while maintaining the psychometric validity of the tool.
- Solid: research by Ree, Earles, and Teachout (1994) demonstrates that for predicting job performance, the assessment of the *g* factor alone is sufficient, and that the addition of specific skills (such as numerical or verbal reasoning) does not provide additional information about a candidate's capacity for success. Therefore, measuring other dimensions, while interesting, unnecessarily extends the length of the test.

3. Theoretical foundations

3.1. Theoretical elements

BRAIN is a psychometric assessment that aligns with the dominant theories of intelligence by relying on the concept of the *g* factor. Following the works of Spearman (1904) and the Cattell-Horn-Carroll (CHC) model (1997), BRAIN measures general intelligence, acknowledged as one of the most significant predictor of professional performance. BRAIN's uniqueness lies in its exclusive focus on the *g* factor, diverging from earlier versions of the test that differentiated various types of reasoning like verbal, analogical, abstract, and numerical. This focus is backed by studies, such as those by Ree, Earles, and Teachout (1994), which have shown that the *g* factor alone provides a sufficient prediction of work performance, making evaluations of specific competencies superfluous. These competencies did not seem to offer significant additional predictive information. Carroll (1993) also supports this approach by indicating that despite the existence of distinct forms of reasoning, they are strongly interrelated and linked to the *g* factor.

Moreover, while detailed analysis of different cognitive abilities may seem beneficial, it significantly increases the time required for testing, which is not proportional to the added predictive value. Therefore, BRAIN is designed to optimise candidates' experience, focusing on conciseness and efficiency. This results in a shorter, more dynamic assessment that respects users' time while providing a reliable measure of general intelligence. This methodological choice represents an advancement in talent assessment, offering companies a streamlined and powerful tool for their selection processes. In line with this streamlined approach, BRAIN is part of a broader strategy to value efficiency in talent assessment. By reducing test duration, it respects candidates' time and promotes better focus and optimal performance during the test. As some studies indicate, when reasoning measures are too lengthy, the results may be influenced by the respondents' personality traits (Myszkowski, Storme, Kubiak, and Baron, 2022). Additionally, the assessment is particularly relevant in a context where the speed and relevance of decision-making are crucial. Ultimately, this *g* factor-centered approach allows for a more seamless integration of the test into human resource management systems, providing more synthetic indicators for analysing and predicting professional performance. Thus, BRAIN positions itself as a strategic tool for companies seeking an effective and concise cognitive evaluation.

3.2. Type of analysis available in BRAIN

Dimension	Description
Reasoning power (global score)	It represents the *g*-factor, or general intelligence, on a scale from 1 to 10, without decimals. A high score suggests a strong learning potential and an increased ability to handle complex tasks, indicating significant professional development potential.
Reasoning speed (global time)	Measured by a gauge surrounding the score, the time indicates how quickly an individual completes the test. A full gauge means maximum use of the allotted time, reflecting a more thoughtful or cautious approach.
Preferred tasks	This dimension assesses the level of complexity of tasks an individual is inclined to handle, ranging from simple routine actions to autonomous intermediate tasks, up to complex and strategic activities.
Decision-making Type	This component measures the decision-making time delay, classifying individuals as quick, reasoned, or cautious based on the time they devote to thinking before deciding.
Learning style	BRAIN identifies how a person learns best by determining whether they tend to innovate, delve deeper, experiment, or apply knowledge and skills.
Behavioral style	Beyond the level of reasoning, BRAIN analyses how responses are given, whether in situations of success or failure, offering insight into underlying behavior.
Strengths	In BRAIN's "Strengths" section, three key elements are scrutinised to understand how a person approaches and processes complex information, makes decisions under pressure, and maintains accuracy in their judgments. Complexity management reveals their ability to grasp nuanced information in decision-making. Decision speed indicates whether they are inclined to make decisions quickly or prefer a more measured pace. Accuracy is an indicator of the correctness and reliability of the decisions made.
Risks factors	BRAIN examines "Risk Factors" that could hinder cognitive performance. Rashness and excessive caution are two extremes that reflect a tendency towards recklessness or over-hesitation in decision-making, respectively. Inaccurate deductions identify a propensity for analytical errors, while indecisiveness is manifested by a reluctance to commit firmly to answers, often indicated by frequent recourse to the "I don't know" option. This balance between strengths and potential risks provides a nuanced assessment of an individual's decision-making capabilities.

Table 3.1. Description of the analysis available in BRAIN.

4. Development of BRAIN

The development of BRAIN proceeded through several phases.

4.1. Phase 0

- Literature review: this step aimed to fully inform the development team about scientific advancements regarding the assessment of reasoning abilities in a professional context, new assessment formats, the impact of gamification on assessment formats, etc.
- Development and testing of a first prototype: acomplete initial prototype (including instructions, items, etc.) was developed and subsequently tested with a sample of AssessFirst employees. This first test provided initial feedback on the understanding of the assessment format, the value of the instructions, the level of the first items, the correlation between scores on the old BRAIN, and performance on this prototype, among other factors.
- Improvement of the prototype: following the internal test phase, the prototype was refined, particularly in terms of instructions and visualisation of the elements present on the game scene. A second test with a different sample of AssessFirst employees was conducted. The changes made to the prototype were conclusive.

4.2. Phase 1

- Launch of the first testing phase: a test version was launched in the first week of October 2019. This version consisted of 20 items and aimed primarily to gather initial feedback from candidates on the assessment and to identify the best items among the 20 proposed (the best items being those that provide information on a candidate's level, allow discrimination between the better and poorer performers, and show a strong correlation with the scores of the old BRAIN). During this phase, which ran from the first week of October 2019 to the second week of November 2019, over 2000 sessions were collected. 6 items, considered the "best," were isolated and used as anchors for the following test phases.

4.3. Phase 2

- Launch of phase 2.1 of testing: based on the analyses of the phase 1 data, five new series of 20 items were created, each incorporating the 6 anchor items identified during phase 1. This phase aimed to analyse the psychometric quality of the newly designed items. During this phase, which ran from the second week of November 2019 to the last week of November 2019, 150 sessions per series were collected, totaling 750 sessions.
- Launch of phase 2.2 of testing: following analyses of phase 2.1 data, five new series of 20 items were created, again incorporating the 6 anchor items from phase 1. This phase continued to analyse the psychometric quality of the items. This phase, which lasted from the last week of November 2019 to the last week of December 2019, resulted in 150 sessions per series, or 750 sessions in total.
- Launch of phase 2.3 of testing: based on the data from phase 2.2, five new series of 20 items were created. These series also included the 6 anchor items from phase 1. This phase aimed to further analyse the psychometric quality of the items. During this phase, which took place from the

last week of December 2019 to the last week of January 2020, 150 sessions per series were collected, for a total of 750 sessions.

- Launch of phase 2.4 of testing: based on the analyses of the phase 2.3 data, five new series of 20 items were created. These series incorporated the 6 anchor items from phase 1. This phase focused on analysing the psychometric quality of the items. During this phase, which ran from the last week of January 2020 to the second week of February 2020, 150 sessions per series were collected, totaling 750 sessions.

4.4. Phase 3

- The launch of phase 3 of testing: based on the analyses of all data from Phase 2, we created 5 new series of 20 items each. These series were composed of the 6 anchor items identified during Phase 1 plus 14 items from Phase 2 considered among the "best." In total, 70 items from the Phase 2 test were selected and integrated into these 5 new series (thus 76 items in total, counting the 6 anchor items). This phase aimed to collect data for the calibration of the assessment. During this phase, which took place between the third week of February 2020 and the first week of April 2020, we collected just over 500 sessions per series, totaling 2800 sessions.

4.5. Phase 4

- Calibration and construction of adaptive models: the results from Phase 3 were analysed to perform calibrations for the different scores and indicators of BRAIN. The adaptive models were also built based on the results and levels of items from Phase 3. This calibration phase occurred between the second week of April 2020 and the last week of April 2020.
- Migration: the adaptive test was then integrated into the application (in simulation and not in production) to replace the old BRAIN. Necessary adjustments for its integration (modifications of reports, some visual changes, migration of old scores, etc.) were made during this phase. It took place between the second week of April 2020 and the second week of May 2020.
- QA testing: the third week of May was dedicated to QA testing.
- Production Launch: the assessment went live on Tuesday, June 9, 2020.

The methodology employed for the development of BRAIN reflects a rigorous and iterative research and development process, characterised by a series of 6 phases of testing and data collection. Each phase was methodically designed to refine the test based on feedback, item performance, and their correlation with previous versions, thus ensuring an evidence-based evolution. The scale of this project is exemplified by the approximately 8000 sessions conducted, which played a critical role in transitioning from the initial to the final version. This comprehensive process, spread over several months, led to a precisely calibrated BRAIN tool, equipped with sophisticated adaptive models, and ready for successful integration into the final application. The implementation of this methodology demonstrates a commitment to scientific excellence and user experience, establishing BRAIN as a state-of-the-art general intelligence assessment tool for the professional environment.

5. Validity

How can we determine if an assessment accurately measures what it claims to measure? How can we ensure that each scale is measured correctly and that the results of the assessment have the intended meaning? These questions are answered through validation studies. The purpose of validating an assessment is to confirm that it measures the intended construct and to determine the accuracy of the results obtained from it. In the past, validity was defined as the correlation between a score on an assessment and an external criterion that measured either the same construct or a construct that was supposed to be related to the construct associated with the score. To establish and ensure the validity of an assessment, several types of validity must be examined. The validity studies of BRAIN cover the following types of validity:

- **Construct validity**: refers to the degree to which the assessment accurately measures the psychological construct or dimension it is designed to assess. This type of validity is established through various analyses, such as RMSEA, and distribution parameters;
- **Convergent validity**: refers to the degree to which two measures of constructs that should theoretically be related are indeed related. In other words, convergent validity measures the degree to which the results of one assessment are correlated with those of another assessment that assesses the same or a similar concept;
- **Predictive validity**: the predictive validity of a personality assessment measures its ability to predict a target variable, such as job performance or turnover. In other words, it assesses whether the results of the personality assessment can be used to predict future outcomes in the workplace.

5.1. Construct validity

Construct validity refers to whether an assessment instrument measures the intended theoretical construct and not something else. It is closely related to other aspects of validity, as any evidence of validity contributes to understanding the construct validity of a test. The importance of construct validity lies in the fact that it influences the interpretation of test scores. If a test claims to measure a specific dimension, it is crucial to ensure that it actually measures that dimension. Otherwise, any interpretation of the scores would be incorrect and could lead to biased decisions. However, construct validity is not limited to simply looking at whether the assessment is measuring a specific dimension. It involves a comprehensive investigation to determine whether the interpretations of the test results are consistent with the theoretical and observational terms that define the construct (Cronbach & Meehl, 1955).

There is no single method for determining construct validity, but rather different methods and approaches must be combined. In order to assess the construct validity of BRAIN, we have utilised two complementary methods: the Root Mean Square Error of Approximation (RMSEA), and the presentation of distribution parameters (Fisher, 1912, 1920, 1921, 1922).

- The **RMSEA**, or Root Mean Square Error of Approximation, measures the difference between the observed data and the fitted data of the model, corrected for the number of free parameters of the model. The RMSEA assesses the absolute fit of the model by comparing the unexplained variance in the data with the expected unexplained variance in the population given the model. Generally, an RMSEA < .05 indicates a good fit of the model to the data (Steiger & Lind, 1980; Browne & Cudeck, 1993);

- **The distribution parameters** correspond to the statistical characteristics of the distribution of scores in the assessments. These parameters include measures such as the mean, standard deviation, skewness, and kurtosis, which provide information about the shape, centre, and variability of the distribution. They make it possible to identify atypical scores, explore individual differences in the distribution of scores, and to better interpret the results. For example, high levels of skewness or kurtosis may indicate non-normal distributions, which could affect the interpretation of the results and the use of certain statistical tests. Therefore, it is important to examine distribution parameters in addition to other aspects of validity when assessing the quality of an assessment.

5.1.1. RMSEA

The latest RMSEA studies for BRAIN were conducted in 2020. For the overall test score, the RMSEA index is below .05 (RMSEA = .04), indicating a good fit of the model to the data. This suggests that the model is effective in explaining the relationships between the measured variables and that the discrepancies between the observed data and the predictions made by the model are minimal. This information provides further evidence of construct validity for the BRAIN assessment.

Dimension	*RMSEA*
Global score	.04

Table 5.1. RMSEA for BRAIN's global score.

5.1.2. Distribution parameters

The distribution of scores obtained from a reasoning assessment is an essential aspect of the assessment's construct validity. The way scores are distributed can provide vital information about how the test measures the global dimension and how the scores are interpreted. Our analysis of distribution parameters focuses on five primary parameters that are necessary:

- the **mean,** which is an indicator of the central tendency of the scores in the distribution;
- the **median**, which is a measure of central tendency that represents the value that divides a distribution in half, with 50% of the scores above and 50% below the median. Unlike the mean, the median is less sensitive to extreme scores and is a more robust measure of central tendency;
- the **standard deviation**, which is a measure of the dispersion of scores around the mean. It is calculated by taking the square root of the variance of the scores;
- **skewness**, which is a measure of the symmetry of a distribution. It is calculated by comparing the frequency of scores to the left and right of the mean. If the distribution is perfectly symmetric, the skewness is zero. If the distribution is skewed to the left, the skewness is negative. If the distribution is skewed to the right, the skewness is positive;
- **kurtosis,** which is a measure of the degree of peakedness or flatness of a distribution compared to a normal distribution. A normal distribution has a kurtosis of 0. If a distribution is more peaked than a normal distribution, its kurtosis value is positive, and if it is less peaked, its kurtosis value is negative.

Expectations for distribution parameters depend on the context and the measurement instrument used. However, in general, here is what is expected for "good" distribution parameters:

- the mean should be close to the median value: this indicates that the distribution is symmetric. If the mean is significantly different from the median, this may indicate an asymmetry in the distribution;

- the standard deviation should be reasonable and large enough to capture individual differences in the measured dimension, but not so large as to dilute the differences between individuals. In general, one would expect the standard deviation to be around 2 for the 10-point personality scale;
- the asymmetry (skewness) should be close to 0 (symmetrical distribution). If the skewness is significantly different from 0, this may indicate an asymmetry in the distribution;
- the kurtosis should be close to 0 (normal distribution). If the kurtosis is significantly different from 0, the distribution is either flatter or more peaked.

The latest studies on the distribution parameters of BRAIN were conducted in 2023 (N = 50,000).

Dimension	Mean	Median	SD	Skewness	Kurtosis
Global score	5.65	6.00	1.90	-0.05	0.18
Global time	49.62	50.00	29.09	0.20	-1.22

Table 5.2. Distribution parameters for the global score and the global time.

The distribution parameters are consistent with the expected standards. The normality of the distributions can be interpreted through indicators such as the alignment of the means and medians, as well as the symmetry and kurtosis coefficients being close to 0.

5.2. Convergent validity

5.2.1. Introduction

Convergent validity is a measure of how similar the scores of a reasoning test are to scores from other tests or measures that assess the same reasoning dimension or factor. This allows us to verify whether a reasoning test accurately measures what it is intended to measure. Specifically, convergent validity is determined by the correlation between the scores of a test and those of other measures or tests that assess the same dimension of reasoning. A strong correlation between the scores indicates that the scales are measuring the same construct, which strengthens the validity of the test.

It should be noted that there is no "official" threshold for judging the quality of convergence between two measures. Additionally, the appropriate threshold depends on the specific context in which the assessment is used and the characteristics of the target population. Furthermore, convergent validity must be assessed in conjunction with other measures of validity to have a complete assessment of the quality of the personality test. However, several authors and researchers have offered some suggestions or satisfaction thresholds: (1) a correlation of .7 or more between an assessment and other measures that assess the same dimension is an indicator of very strong convergent validity, according to Campbell and Fiske (1959); (2) a correlation of .6 is recommended as a threshold of validity by Worthington and Whittaker (2006); (3) a correlation of .5 or more is considered good convergent validity by Bagozzi and Yi (1988) and by Revelle and Condon (2015); (4) a correlation of .4 is considered acceptable by Nunnally and Bernstein (1994). In short, although there is no clear consensus or golden rule (Marsh, Hau & Wen, 2004) on the exact value to use as the threshold of convergent validity, it is recommended to aim for correlations of .5 or higher to support good convergent validity of a personality assessment.

5.2.2. Convergent validity with IMak-17

As part of the development of BRAIN, we are studying the convergent validity of the assessment with a matrices test specifically generated for our study. This psychometric test, inspired by Raven's matrices, is used to measure fluid intelligence, which is the ability to reason and solve new problems without relying on previous knowledge. The test consists of a series of matrices or visual patterns where the participant must identify the missing part to complete the matrix. It is often used in educational and professional contexts to assess logical thinking and visual analysis skills. This study was conducted in 2021.

5.2.2.1. Participants and procedure

For this study, participants registered on the AssessFirst application were recruited. An email with a link to the study was sent to the last 11,000 users of the application who had previously completed BRAIN and had agreed to be contacted for scientific studies. The email specified that this survey was related to the BRAIN test they had already taken on the platform, but that their responses and results would be used solely for scientific purposes, without impacting their online assessment profile or being shared with external companies. Therefore, this study represents a low-stakes context. 7.4% of users responded voluntarily. Furthermore, as the matrices test was not optimised for display on smartphone screens, we excluded participants who completed the survey on a smartphone and only retained those who did so on a computer. In this final sample (N=555), 53.5% identified as women and 46.5% as men, with an average age of 38.9 years. All conditions and data exclusions are reported in this document. Participants were contacted based on their prior response to the BRAIN test, though some may have taken other measures on the AssessFirst platform; however, these measures were not used in this study.

Once connected, participants took the matrices test, which consisted of 17 items. The time taken to respond to each item was recorded, as well as the selected response option. Response times were recorded directly via the survey application used (SurveyGizmo), by calculating the difference (in seconds) between the item display time and when the participant clicked to submit their answer. At the end of the study, participant responses were linked to AssessFirst application data, which allowed us to use the previously completed BRAIN assessment and to study the convergent validity between the two tests.

5.2.2.2. Measure

We used the IMak library (Blum & Holling, 2018) to generate analogies in the form of matrices to assess intelligence. The final test consisted of 4 training items on a rule, 4 items on one rule, 6 items on two rules, 6 items on three rules, and 1 item on four rules, presented in this order. In this study, this test will be referred to as IMak-17. Since the number of rules is an indicator of item difficulty (Blum & Holling, 2018), this test can be considered as a progressive matrices test. We estimated its reliability using McDonald's omega (Flora, 2020), which was $\omega = 0.84$.

5.2.2.3. Objective

In our study, we aimed to test the convergent validity of BRAIN by comparing it to IMak-17. Convergent validity is a type of validity that measures to what extent two independent tests, supposed to assess the same psychological construct, correlate. For this, we use statistical correlation analysis, which assesses the degree of linear relationship between the scores obtained on the two tests. If we observe a strong correlation between BRAIN and IMak-17, this would indicate that both tests measure similar aspects of intelligence, suggesting strong convergent validity. Conversely, a weak correlation would suggest that the tests may not be measuring the same construct or that one is not a good indicator of intelligence.

The strength of the correlation is generally measured using Pearson's correlation coefficient (r). The values of this coefficient range from -1 to +1, where +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. Generally, an r value beyond 0.70 is considered strong, around 0.50 is considered moderate, and below 0.30 is considered weak. For our analysis, we collected participants' scores on BRAIN and compared them with their scores on IMak-17. We then calculate the Pearson correlation coefficient to assess the strength of the relationship between the two sets of scores. If a strong correlation is found, this will allow us to conclude with more confidence that BRAIN has good convergent validity in terms of measuring intelligence, similar to IMak-17.

It is important to note that the results may be influenced by the context of the test administration. Indeed, the correlation could probably be attenuated by the time elapsed between the two measures, differences in item content (especially due to the adaptive and gamified aspect of the BRAIN test, based on CATs), and, more importantly, the fact that IMak-17 was completed by participants in low-stakes conditions, whereas the BRAIN test was completed under high-stakes conditions (since the results of the BRAIN test are used to feed the person's online profile in the application).

5.2.2.4. Results

The analyses report a Pearson correlation coefficient of $r = 0.74$. This significant correlation of 0.74 suggests a strong positive relationship between the scores obtained on BRAIN and those on IMak-17. In terms of convergent validity, this indicates that BRAIN is a robust indicator of intelligence, similar to IMak-17. A coefficient of this magnitude allows us to conclude that both tests evaluate the same psychological construct, that is intelligence, with a high degree of similarity.

Furthermore, such a high correlation also suggests that variations in participant performance on one test can be reliably predicted by their performance on the other test. This reinforces the credibility of BRAIN as a cognitive assessment tool and justifies its use in contexts where an accurate measure of intelligence is necessary. It is important to note that while the correlation is strong, it is not perfect. This indicates that although the tests share significant common variance, there are also differences that could be attributed to the specificities of each test or to other factors measuring slightly different aspects of intelligence. Nevertheless, a correlation of 0.74 is considered solid evidence of convergent validity.

5.3. Predictive validity

Predictive validity of a motivation assessment is the measure of its ability to predict a target variable, such as performance or turnover. In other words, it is about whether the assessment results can be used to forecast future performance. Evidence of predictive validity is particularly relevant when one wishes to infer, from a assessment score, an individual's position on another criterion variable that is independently assessed at a later date. Studies of predictive validity related to BRAIN are presented in a dedicated guide.

5.4. Conclusion

Validating a reasoning assessment is essential to ensure that the obtained measurements are precise. In this study, we examined the construct validity, convergent validity, and predictive validity of BRAIN. Our analyses indicate that: (1) the assessment is well-structured and demonstrates good measurement homogeneity, (2) BRAIN has strong convergent validity with IMak-17, and (3) BRAIN is predictive of job performance. Overall, the results meet the most demanding psychometric standards and demonstrate the validity of BRAIN. However, to further delve into the psychometric qualities of BRAIN, it is necessary to consider its reliability. In this sense, a assessment must be both valid and reliable to be used in professional decision-making contexts (recruitment, mobility, etc.). Indeed, a valid but not reliable assessment would mean that the test accurately measures what it is supposed to measure, but individual scores are inconsistent. On the contrary, a assessment that is both valid and reliable consistently measures what it is supposed to measure—in other words, it consistently hits the bullseye. Evidence of reliability is presented in the following chapter.

6. Reliability

How can you determine if the results of an assessment are reliable? How can you ensure that the assessment produces consistent results when asking the same questions to the same person at different times? The answers to these questions can be obtained through the study of reliability. Whilst validity provides information on an assessment's ability to measure what it intends to measure, reliability measures whether the measurement is consistent and reliable every time the same assessment is completed by the same person. In short, the reliability of an assessment measures its consistency or stability over time and aims to determine if an assessment produces similar results when asking the same questions to the same person at different times or to similar people. Therefore, the objective of reliability is to ensure that the obtained results are dependable and accurate. BRAIN's reliability is analysed here through test-retest reliability. This method evaluates the reliability of a measurement by assessing the same variable at two different times. It measures the temporal stability of the measure and estimates the proportion of total variance attributable to measurement error. It is often used in longitudinal studies or to assess the stability of a test over a given time period. This involves administering the same assessment to a group of participants at two different times, with a time interval between the two sessions. The correlation between the results is calculated to determine the reliability. If this correlation is high, it means that the scores are stable, and thus the test can be considered reliable.

The latest test-retest reliability studies for BRAIN were conducted in 2023 with a sample size of 2,309 participants. The two sessions were administered with a minimum interval of three months between them. The overall score for the BRAIN assessment (general reasoning ability) shows good test-retest reliability with a coefficient $r = .81$. This stability of scores over a significant period highlights BRAIN's ability to provide consistent and reliable measures of an individual's level of reasoning, which is an essential asset for assessments in a professional context where long-term decisions are based on these data.

Dimension	*Pearson r*
Global score	.81

Table 6.1. Test-retest reliability for BRAIN's global score.

7. Sensitivity

Sensitivity, also called discrimination, refers to the ability of an assessment to distinguish between people with a high level on a dimension and people with a low level. It, therefore, reflects the ability of the assessment to identify the uniqueness of each individual. A sensitive reasoning assessment can identify the subtle differences between people, and help to understand their reasoning skills with more precision and discrimination. Sensitivity thus refers to the ability of the assessment to correctly identify people who possess the characteristic being measured and to avoid false positives (people who do not possess the characteristic, but who are identified as such by the assessment).

In the development of BRAIN, we are studying the discriminant power and sensitivity of each test item using the alpha parameter (α). This discrimination parameter is an integral part of Item Response Theory (IRT) models, which is a modern approach to psychometric test design. The discrimination parameter, α, reflects an item's ability to differentiate among respondents based on their level of the ability or trait being measured. An item with a high discrimination parameter is more effective in distinguishing between individuals who have different levels of the dimension in question. For example, in a math competence test, an item with high discrimination will better differentiate between individuals with high and low mathematical competence. In practice, discrimination parameters are estimated during item calibration and can be used to select the most informative items for a test. They play a particularly important role in the construction of adaptive tests (CAT), where items are dynamically chosen to match the estimated competence level of the respondent.

The α parameter is a crucial indicator in assessing an item's discrimination in psychometric tests, serving to measure its ability to differentiate individuals based on their level of ability. According to Item Response Theory, the value of the α parameter has a direct relationship with the item's discriminant power: the higher the value of α, the better the discrimination. To facilitate the interpretation of α values, Baker (2001) introduced a normative evaluation grid that categorises the discriminative capacity of items as follows:

- Discrimination is considered "null" if α equals 0, meaning the item does not differentiate between respondents' ability levels.
- It is deemed "very low" if α is between 0.01 and 0.34, indicating the item has minimal discrimination capacity.
- Discrimination is "low" for α values ranging from 0.35 to 0.64, showing limited ability to distinguish respondents based on their aptitude.
- An α value between 0.65 and 1.34 corresponds to "good" discrimination, indicating good efficiency for separating ability levels.
- Discrimination is qualified as "strong" if α lies within the range of 1.35 to 1.69, denoting a superior ability to differentiate individuals effectively.
- Finally, discrimination is considered "very strong" when α exceeds 1.70, meaning the item is extremely effective at identifying capability differences among respondents.

This grid is a valuable tool in the design and evaluation of tests, as it provides clear benchmarks to judge the quality and usefulness of items within a measurement scale.

The latest sensitivity studies for BRAIN, based on the calculation of the α parameter, were conducted in 2020. The evaluation of the discriminative power of the test items reveals a strong capacity to distinguish individuals based on their aptitude. This evaluation is based on the distribution of the discrimination values of individual items, categorised according to Baker's (2001) evaluation grid:

- It is encouraging to note that none of the BRAIN items show "null" discrimination, meaning there are no items incapable of differentiating respondents based on their aptitude.
- None of the BRAIN items show "very low" discrimination
- A small portion of the inventory, specifically 6 items, displays "low" discrimination. Although these items contribute less effectively to differentiating the level of aptitude, they can still provide useful information in the overall context of the test.
- The majority of the items, totaling 31, are classified as having "good" discrimination. These items are capable of effectively distinguishing between individuals at different ability levels and form the solid backbone of the test.
- 16 items have been identified with "strong" discrimination, reflecting their excellent ability to accurately differentiate between people with varied levels of the trait.

- Finally, a notable number of 23 items demonstrate "very strong" discrimination. These items are particularly powerful and effective at separating respondents according to their level of aptitude, thus contributing significantly to the overall sensitivity of the test.

In summary, the BRAIN test is characterised by a high psychometric quality, with a predominance of items offering discrimination ranging from good to very strong. The inclusion of items with low discrimination in the BRAIN test is a deliberate choice that enriches the test's dimensionality and its ability to assess the entire spectrum of abilities. These items, although less discriminating in the general context, are valuable because they provide specific information at the extremes of the difficulty spectrum. Low discrimination items are particularly useful for assessing individuals with very low or very high abilities. For respondents with very low skills, these items may be among the only ones they can answer correctly, thus providing a measure of ability where more discriminating items could not. At the other end, for individuals with very high abilities, these items can help to identify the threshold at which they begin to experience difficulty. Therefore, even though these low discrimination items contribute less to differentiating respondents in the middle of the ability spectrum, they are essential for capturing variability at both extremes. This maintains a balanced test, which can provide relevant measures for all skill levels, thus ensuring a comprehensive and nuanced assessment of intelligence.

Item	α	Item	α	Item	α	Item	α
item_id_001	0.53	item_id_020	2.03	item_id_039	1.24	item_id_058	2.76
item_id_002	1.72	item_id_021	1.25	item_id_040	1.95	item_id_059	1.88
item_id_003	1.47	item_id_022	0.82	item_id_041	0.49	item_id_060	2.38
item_id_004	1.08	item_id_023	1.94	item_id_042	2.27	item_id_061	1.61
item_id_005	0.82	item_id_024	1.63	item_id_043	1.45	item_id_062	1.23
item_id_006	1.60	item_id_025	1.45	item_id_044	1.57	item_id_063	1.68
item_id_007	0.88	item_id_026	1.98	item_id_045	1.20	item_id_064	0.44
item_id_008	0.70	item_id_027	0.84	item_id_046	1.01	item_id_065	0.79
item_id_009	1.44	item_id_028	0.85	item_id_047	1.64	item_id_066	1.29
item_id_010	0.72	item_id_029	1.90	item_id_048	1.71	item_id_067	0.71
item_id_011	0.46	item_id_030	2.46	item_id_049	1.41	item_id_068	0.70
item_id_012	1.28	item_id_031	1.10	item_id_050	0.55	item_id_069	3.60
item_id_013	1.37	item_id_032	0.70	item_id_051	1.91	item_id_070	0.90
item_id_014	2.52	item_id_033	1.70	item_id_052	1.28	item_id_071	1.26
item_id_015	1.58	item_id_034	1.65	item_id_053	1.26	item_id_072	1.29
item_id_016	1.77	item_id_035	0.92	item_id_054	1.46	item_id_073	1.22
item_id_017	0.53	item_id_036	1.71	item_id_055	2.72	item_id_074	1.84
item_id_018	0.95	item_id_037	0.77	item_id_056	1.20	item_id_075	1.75
item_id_019	1.82	item_id_038	1.42	item_id_057	0.77	item_id_076	1.81

Table 7.1. Sensitivity of BRAIN's items.

The descriptive statistics regarding the discrimination power of the BRAIN test items highlight a general trend towards high discrimination. The average and median values, very close to each other, stand at 1.40 and 1.39 respectively, indicating strong discrimination according to Baker's (2001) classification. This proximity between the mean and median also suggests a relatively symmetrical distribution of discrimination values around a high level, which is characteristic of a well-calibrated test. The observed minimum of 0.44 falls into the category of low discrimination. While this is the lowest discrimination value among the test items, it does not reach the threshold of very low or null discrimination, meaning that even the least discriminating item still adds some value to the test. Conversely, the recorded maximum of 3.60 is well beyond the threshold for very strong discrimination. This extreme value reflects a particularly powerful item that can very effectively distinguish participants according to their ability levels. These results demonstrate that the BRAIN test, as a whole, has a strong discriminatory capacity, with a distribution of discrimination values that favors an accurate assessment of cognitive abilities across a wide range. The presence of items with both low and very high discrimination values allows the test to effectively cover the entire difficulty spectrum, thereby ensuring a comprehensive measure of intelligence across various levels of competency.

8. Fairness

Fairness in the context of a reasoning assessment refers to the extent to which it is designed to be fair and unbiased for all individuals, regardless of their origin, gender, sexual orientation, race, or culture. In other words, a fair assessment should be objective and impartial towards all individuals who take it, without any bias or discrimination against any particular group. Our teams take every measure to ensure the fairness of our assessments and predictive analyses, and we ensure that the use of our algorithms in decision-making processes does not lead to discrimination through any unforeseen algorithmic biases. Additionally, in the development of our assessments, equity studies focus on two areas: (1) ensuring the accessibility of the assessment, and (2) ensuring equity in the results of the assessment.

8.1. Fairness in BRAIN's results

The data presented in this section highlights that the results of BRAIN do not show significant differences or strong effect sizes based on gender and age variables. It is important to note that AssessFirst only requests personal information necessary for the appropriate use of the platform. For instance, we do not collect information about religious, political, or sexual orientation. Regarding age, we only ask for date of birth to ensure it does not impact how questions are handled. Moreover, the variables analysed below do not play any role in the calculation of results within the AssessFirst solution. Our commitment to protecting user privacy and promoting inclusivity is reflected in our data practices.

8.1.1. Fairness regarding gender

The latest gender equity analyses for BRAIN were conducted in 2022, with a sample (N = 332,587) composed of 51% men and 49% women. Overall, the average scores for the overall BRAIN score for both men and women are between 5 and 6, which is close to the theoretical average of 5.5. Therefore, there is no major difference between the results of men and women on the overall score measured by the BRAIN assessment. The results are thus equitable according to the gender of the respondent.

Dimension	Male	Female
Global score	5.74	5.59

Table 8.1. Mean global score depending on gender.

Moreover, the effect size highlighted here supports the previous conclusion. Indeed, Cohen's d is equal to .07, which demonstrates the absence of a difference between the results of men and those of women.

Dimension	Cohen's d	Effect size
Global score	.07	-

Table 8.2. Mean global score depending on gender.

8.1.2. Fairness regarding age

The latest age equity analyses for BRAIN were conducted in 2022, on a sample (N = 64,310) consisting of 52% of individuals over 35 years old and 48% under 35 years old. Overall, the average scores for the overall BRAIN score for both age groups are between 5 and 6, close to the theoretical average of 5.5. Therefore, there is no major difference between the results of individuals over 35 years old and those under 35 years old on the overall score measured by the BRAIN assessment. The results are thus equitable according to the age of the respondent..

Dimension	≥ 35 years old	< 35 years old
Global score	5.45	5.67

Table 8.3. Mean global score depending on age.

The effect size here also supports the previous conclusion. Cohen's d is equal to -.11, which demonstrates the absence of a difference between the results of individuals over 35 years old and those under 35 years old on the overall score measured by the BRAIN assessment.

Dimension	Cohen's d	Effect size
Global score	-.11	-

Table 8.4. Mean global score depending on age.

8.1.3. Conclusion

Indeed, whether in terms of gender or age categories, there are no major differences in the results obtained by different groups on the BRAIN assessment from AssessFirst. In summary, the results support the hypothesis that the assessment does not discriminate against a particular gender or age.

Conclusion

The psychometric studies presented in this technical manual demonstrate and attest to the scientific robustness of the assessments developed by AssessFirst. The various analyses show the validity, reliability, sensitivity, and fairness of each assessment. It's crucial to highlight that these results were achieved through a rigorous process of development and validation of the assessments, adhering to the strictest international standards in psychometry. Compliance of these tools with the standards recommended by the American Psychological Association (APA) and the International Test Commission (ITC) allows AssessFirst to guarantee a high level of quality in the design of assessments and to continuously improve the reliability of its assessment tools. These efforts and commitment to quality meet the requirements of human resources professionals in evaluating candidates and employees.

Further analyses will be regularly added to this manual to perfect the demonstration of scientific robustness. Also, the roadmap for upcoming studies includes: (1) the predictive validity of SWIPE - between May and the end of 2023, (2) the test-retest reliability of SWIPE in January 2024, (3) studies related to the fairness regarding hierarchical status for SWIPE.

For more information on the scientific aspects related to our tools and product, you can contact your Account Manager and/or Customer Success representative, or one of our below experts.

Emeric KUBIAK
Psychologist
Head of Science @AssessFirst

Simon BARON
Psychologist
Chief Product Officer @AssessFirst

References

Allport, G. W. (1961). Pattern and growth in personality. Holt, Reinhart & Winston.

Ames, D. R., Kammrath, L. K., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. Personality and Social Psychology Bulletin, 36(2), 264–277. doi: 10.1177/0146167209354519

Anastasi, A. (1954). Psychological testing. Macmillan Co.

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. Journal of Applied Psychology, 76(5), 732-740. doi: 10.1037/0021-9010.76.5.732

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B., & Landers, R. N. (2016). Correcting misconceptions about gamification of assessment: More than SJTs and badges. Industrial and Organizational Psychology: Perspectives on Science and Practice, 9(3), 671–677. doi: 10.1017/iop.2016.69

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. American Psychologist, 55(5), 469–480. doi: 10.1037/0003-066X.55.5.469

Arthur, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The Use of Mobile Devices in High-stakes Remotely Delivered Assessments and Testing. International Journal of Selection & Assessment, 22(2), 113-123. doi:10.1111/ijsa.12062

Arthur, W., & Traylor, Z. (2019). Mobile Assessment in Personnel Testing: Theoretical and Practical Implications. In R. Landers (Ed.), The Cambridge Handbook of Technology and Employee Behavior (Cambridge Handbooks in Psychology, pp. 179-207). Cambridge: Cambridge University Press. doi:10.1017/9781108649636.009

Ashton, M. C., & Lee, K. (2019). How Well Do Big Five Measures Capture HEXACO Scale Variance ? Journal of Personality Assessment, 101(6), 567-573. doi: 10.1080/00223891.2018.1448986

Ashton, M. C., Lee, K., & Visser, B. A. (2019). Where's the H ? Relations between BFI-2 and HEXACO-60 scales. Personality and Individual Differences, 137, 71-75. doi: 10.1016/j.paid.2018.08.013

Bagozzi, R. & Yi, Y. (1988). On the evaluation of structural equation models. Journal of the Academy of Marketing Science, 16, 74-94. doi: 10.1007/BF02723327

Baron, S., Storme, M., Myszkowski, N., & Kubiak, E. (2023). Forced-choice items: when the respondent cannot choose. 2023 European Congress of Psychology, Brighton, UK.

Bartram, D., & Brown, A. L. (2004). Online Testing: Mode of Administration and the Stability of OPQ 32i Scores. International Journal of Selection and Assessment, 12(3), 278-284. doi: 10.1111/j.0965-075x.2004.282_1.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. doi: 10.1111/j.1744-6570.1991.tb00688.x

Benson, A., Li, D., & Shue, K. (2022). Potential and the gender promotion gap.

Benton, T. (2013). An empirical assessment of Guttman's Lambda 4 reliability coefficient. International Meeting of the Psychometric Society, Arnhem, July 2023.

Berge, J. M. F. T., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69(4), 613-625. doi: 10.1007/bf02289858

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. Psychological Bulletin, 148(7-8), 588–619. doi: 10.1037/bul0000365

Böhm, S., & Jäger, W. (2016). Mobile Candidate Experience: Anforderungen an eine effiziente Bewerberansprache über mobile Karriere-Websites. HMD Praxis der Wirtschaftsinformatik, 53(6), 785-801. doi: 10.1365/s40702-016-0270-5

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons. doi: 10.1002/9781118619179

Bourque, J., Doucet, D. R., LeBlanc, J., Dupuis, J. B., & Nadeau, J. (2019). L'alpha de Cronbach est l'un des pires estimateurs de la consistance interne: une étude de simulation. Revue des sciences de l'éducation, 45(2), 78-99. doi: 10.7202/1067534ar

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. Journal of Personality, 54(1), 106–148. doi: 10.1111/j.1467-6494.1986.tb00391.x

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Sage. doi: 10.4236/jmp.2013.41019

Buckley, M. R., Norris, A. E. W., & Wiese, D. S. (2000). A brief history of the selection interview: may the next 100 years be more fruitful. Journal of management history, 6(3), 113-126. doi: 10.1108/eum0000000005329

Burisch, M. (1997). Test length and validity revisited. European Journal of Personality, 11(4), 303-315. doi: 10.1002/(sici)1099-0984(199711)11:4

Callender J., & Osburn H. (1977). A Method for Maximizing and Cross-Validating Split-Half Reliability Coefficients. Educational and Psychological Measurement, 37, 819-826.

Callender J., & Osburn H. (1979). An Empirical Comparison of Coefficient Alpha, Guttman's Lambda2 and Msplit Maximized Split-Half Reliability Estimates. Journal of Educational Measurement, 16, 89-99. doi: 10.1111/j.1745-3984.1979.tb00090.x

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

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. Journal of Applied Psychology, 101(7), 958-975. doi: 10.1037/apl0000108

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. Journal of Applied Psychology, 104(11), 1347–1368. doi: 10.1037/apl0000414

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi: 10.18637/jss.v048.i06

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New science of talent prediction: Analytics, assessment, and performance. Current Opinion in Behavioral Sciences, 10, 97-101. doi: 10.1016/j.cobeha.2016.04.004

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. Psychological Methods. Advance online publication. doi: 10.1037/met0000475

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. doi: 10.1037/1040-3590.7.3.309

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. Journal of Applied Psychology, 104(10), 1243-1265. doi: 10.1037/apl0000406

Cortina, J. M. (1993). What is coefficient alpha ? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98-104. doi: 10.1037/0021-9010.78.1.98

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. Journal of Personality and Social Psychology, 81(2), 322–331. doi: 10.1037/0022-3514.81.2.322

Courtois, R., Petot, J., Lignier, B., Lecocq, G., & Plaisant, O. (2018). Le Big Five Inventory français permet-il d'évaluer des facettes en plus des cinq grands facteurs ? L'Encéphale, 44(3), 208-214. doi: 10.1016/j.encep.2017.02.004

Courtois, R., Petot, J., Plaisant, O., Allibe, B., Lignier, B., Réveillère, C., Lecocq, G., & John, O. P. (2020). Validation française du Big Five Inventory à 10 items (BFI-10). L'Encéphale, 46(6), 455-462. doi: 10.1016/j.encep.2020.02.006

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334 (1951). doi: 10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. doi.org/10.1037/h0040957

Dalal, D. K., Zhu, X. (S.), Rangel, B., Boyce, A. S., & Lobene, E. (2021). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. Journal of Business and Psychology, 36(1), 55–70. doi: 10.1007/s10869-019-09655-6

David, G., & Cambre, C. (2016). Screened Intimacies: Tinder and the Swipe Logic. Social media and society, 2(2), 205630511664197. doi: 10.1177/2056305116641976

Denissen, J. J. A., Soto, C., Geenen, R., John, O. P., & Van Aken, M. A. G. (2021). Incorporating prosocial vs. antisocial trait content in Big Five measurement: Lessons from the Big Five Inventory-2 (BFI-2). Journal of Research in Personality, 96, 104147. doi: 10.1016/j.jrp.2021.104147

DeYoung, C. G., Carey, B. T., Krueger, R. F., & Ross, S. E. (2016). Ten aspects of the Big Five in the Personality Inventory for DSM–5. Personality Disorders: Theory, Research, and Treatment, 7(2), 113-123. doi: 10.1037/per0000170

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. Journal of Personality and Social Psychology, 93(5), 880–896. doi: 10.1037/0022-3514.93.5.880

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. Annual Review of Psychology, 41(1), 417-440. doi: 10.1146/annurev.ps.41.020190.002221

Dou, X., & Sundar, S. S. (2016). Power of the Swipe: Why Mobile Websites Should Add Horizontal Swiping to Tapping, Clicking, and Scrolling Interaction Techniques. International Journal of Human-computer Interaction, 32(4), 352-362. doi: 10.1080/10447318.2016.1147902

Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014 Aug;105(3):399-412. Epub 2013 Aug 6. doi: 10.1111/bjop.12046

Eagly, A. H., & Revelle, W. (2022). Understanding the magnitude of psychological differences between women and men requires seeing the forest and the trees. Perspectives on Psychological Science, 17(5), 1339–1358. doi: 10.1177/17456916211046006

Efremova, M., Kubiak, E., & Baron, S. (2023). Further understanding of user experience during image-based personality assessment. 2023 European Congress of Psychology, Brighton, UK.

Efremova, M., Kubiak, E., Baron, S., & Frasca, K. (in press). Gender equity in organisational selection: examining the effectiveness of a novel hiring algorithm.

Ferguson, G. A. (1949). On the theory of test discrimination. Psychometrika, 14, 61-68. doi: 10.1007/BF02290141

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. Messenger of Mathematics, 41, 155-160.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Monthly Notices of the Royal Astronomical Society, 80, 758-770.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1, 3-32.

Fisher, R. A. (1922a). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 222, 309-368.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. Applied Psychological Measurement, 5(1), 105-112. doi: 10.1177/014662168100500115

Føllesdal, H., & Soto, C. J. (2022). The Norwegian Adaptation of the Big Five Inventory-2. Frontiers in Psychology, 13. doi: 0.3389/fpsyg.2022.858920

Fyffe, S., Lee, P., & Kaplan, S. (2023). "Transforming" Personality Scale Development: Illustrating the Potential of State-of-the-Art Natural Language Processing. Organizational Research Methods, 109442812311557. doi: 10.1177/10944281231155771

Gallardo-Pujol, D., Rouco, V., Cortijos-Bernabeu, A., Oceja, L., Soto, C. J., & John, O. P. (2021). Factor structure, gender invariance, measurement properties and short forms of the Spanish adaptation of the Big Five Inventory-2 (BFI-2). PsyArXiv. doi: 10.31234/osf.io/nxr4q

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

Gnambs, T. (2016). Sociodemographic effects on the test-retest reliability of the Big Five Inventory. European Journal of Psychological Assessment, 32(4), 307–311. doi: 10.1027/1015-5759/a000259

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

Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48(1), 26–34. doi: 10.1037/0003-066X.48.1.26

Green, S. B. et Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. Psychometrika, 74(1), 121-135. doi: 10.1007/s11336-008-9098-4

Guilford, J. P. (1936). Psychometric methods. McGraw-Hill.

Gutierrez, S.L. & Meyer, J.M. (2013). Assessments on the Go: Applicant Reactions to Mobile Testing. In N.A. Morelli (Chair), Mobile Devices in Talent Assessment: Where Are We Now? Symposium at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9: 139–150. doi: 10.2307/2086306Return

Guttman, L. (1945). A Basis for Analyzing Test-Retest Reliability. Psychometrika, 10, 255-282. doi: 10.1007/BF02288892

Halama, P., Kohút, M., Soto, C. J., & John, O. P. (2020). Slovak adaptation of the Big Five Inventory (BFI-2): Psychometric properties and initial validation. Studia Psychologica, 62(1), 74–87. doi: 10.31577/sp.2020.01.792

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). Multivariate data analysis. Prentice Hall.

Hankins, M. How discriminating are discriminative instruments?. Health Qual Life Outcomes 6, 36 (2008). doi: 10.1186/1477-7525-6-36

Hardy, J. H., Gibson, C., Sloan, M., & Carr, A. (2017). Are applicants more likely to quit longer assessments ? Examining the effect of assessment length on applicant attrition behavior. Journal of Applied Psychology, 102(7), 1148-1158. doi: 10.1037/apl0000213

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. doi: 10.1111/j.1744-6570.2004.00003.x

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. ArXiv:2006.03654 [Cs]. http://arxiv.org/abs/2006.03654

Hilliard, A., Kazim, E., Alatalo, K., & Leutner, F. (2022a). Measuring Personality through Images: Validating a Forced-Choice Image-Based Assessment of the Big Five Personality Traits. Journal of Intelligence, 10(1), 12. doi: 10.3390/jintelligence10010012

Hilliard, A., Kazim, E., Alatalo, K., & Leutner, F. (2022b). Scoring a forced-choice image-based assessment of personality: A comparison of machine learning, regression, and summative approaches. Acta Psychologica, 228, 103659. doi: 10.1016/j.actpsy.2022.103659

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. Journal of Personality and Social Psychology, 93(2), 298–319. doi: 10.1037/0022-3514.93.2.298

Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. Industrial and Organizational Psychology, 1(3), 333-342. doi: 10.1111/j.1754-9434.2008.00058.x

Hill, C. E., Thompson, B. J., & Williams, E. N. (1997). A guide to conducting consensual qualitative research. The Counseling Psychologist, 25(4), 517–572. doi: 10.1177/0011000097254001

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in Hiring. Quarterly Journal of Economics. doi: 10.3386/w21709

Hofmans, J., Kuppens, P., & Allik, J. (2008). Is short in length short in content? An examination of the domain representation in the International Personality Item Pool. Personality and Individual Differences, 45(7), 542-547. doi: 10.1016/j.paid.2008.06.008

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. Psychometrika, 87(2), 749-772. doi: 10.1007/s11336-021-09823-9

Howard, M. C., & Van Zandt, E. C. (2020). The discriminant validity of honesty-humility: A meta-analysis of the HEXACO, Big Five, and Dark Triad. Journal of Research in Personality, 87, Article 103982. doi: 10.1016/j.jrp.2020.103982

Hunt, T. C., & Bentler, P. M. (2015). Quantile Lower Bounds to Reliability Based on Locally Optimal Splits. Psychometrika, 80(1), 182-195. doi: 10.1007/s11336-013-9393-6

Hyde, J. S. (2005). The gender similarities hypothesis. American Psychologist, 60(6), 581–592. doi: 10.1037/0003-066X.60.6.581

Jiao, H., & Lissitz, R. W. (2020). Application of artificial intelligence to assessment. IAP.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), Handbook of personality: Theory and research (pp. 66–100). The Guilford Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big-Five Inventory-Version 4a and 54. Berkeley, CA: Berkeley Institute of Personality and Social Research, University of California. doi: 10.4236/jss.2017.59019

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (pp. 114–158). The Guilford Press.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. Personnel Psychology, 52(3), 621–652. doi: 10.1111/j.1744-6570.1999.tb00174.x

Judge, T. A., & Zapata, C. P. (2015). The person-situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. Academy of Management Journal, 58(4), 1149-1179. doi: 10.5465/amj.2010.0837

Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the five factor model of personality in the large public (N = 320,128). Personality and Individual Differences, 129, 126–130. doi: 10.1016/j.paid.2018.03.026

Kaufman, S. B., Yaden, D. B., Hyde, E., & Tsukayama, E. (2019). The light vs. Dark Triad of personality: Contrasting two very different profiles of human nature. Frontiers in Psychology, 10, Article 467. doi: 10.3389/fpsyg.2019.00467

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. Psychological Methods, 21(1), 69-92. doi: 10.1037/a0040086

Kessler, J. B., Low, C., & Sullivan, C. E. (2019). Incentivized Resume Rating: Eliciting Employer Preferences without Deception. The American Economic Review, 109(11), 3713-3744. doi: 10.1257/aer.20181714

Kim, S. H., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. Asia Pacific Education Review, 11(2), 179-188. doi: 10.1007/s12564-009-9062-8

Kinney, T.B., Lawrence, A. D., & Chang, L. (2014). Understanding the mobile candidate experience: reactions across device and industry. In Kantrowitz & Reddock (chairs) Shaping the Future of Mobile Assessment: Research and Practice Up-date. Symposium at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Kirkebøen, G., & Nordbye, G. H. H. (2017). Intuitive choices lead to intensified positive emotions: An overlooked reason for “intuition bias”? Frontiers in Psychology, 8, Article 1942. doi: 10.3389/fpsyg.2017.01942

Kline, P. (2000). The handbook of psychological testing (2nd ed.). Routledge.

Krainikovsky, S., Melnikov, M., & Samarev, R. (2019). Estimation of psychometric data based on image preferences.Conference Proceedings for Education and Humanities, WestEastInstitute 2019: 75–82.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage.

Kubiak, E., Bernard, B., & Baron, S. (2023). Response speed trajectories as clues of personality in image-based assessment. 2023 European Congress of Psychology, Brighton, UK.

Kubiak, E., Niesner., V., & Baron, S. (2023). Swipe on your personality: measuring facets in 5 minutes through images. 2023 European Congress of Psychology, Brighton, UK.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98(6), 1060–1072. doi: 10.1037/a0034156

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. doi: 10.1016/j.paid.2010.03.042

Lawrence, A. D., & Kinney, T. B. (2017). Mobile devices and selection [white paper]. Society for Industrial and Organizational Psychology.

Lee, K., & Ashton, M. C. (2019). Not much H in the Big Five Aspect Scales: Relations between BFAS and HEXACO-PI-R scales. Personality and Individual Differences, 144, 164-167. doi: 10.1016/j.paid.2019.03.010

Lee, K., Ashton, M. C., & De Vries, R. E. (2022). Examining the expanded Agreeableness scale of the BFI-2. Personality and Individual Differences, 195, 111694. doi: 10.1016/j.paid.2022.111694

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from “human writing” to “machine generation” in personality test development: An application of state-of-the-art natural language processing. Journal of Business and Psychology, 38(1), 163-190. doi: 10.1007/s10869-022-09864-6

Leutner, F., Akhtar, R., & Chamorro-Premuzic, T. (2022). The Future of Recruitment. Emerald Publishing Limited eBooks. doi: 10.1108/9781838675592

Leutner F, Chamorro-Premuzic T. Stronger Together: Personality, Intelligence and the Assessment of Career Potential. J Intell. 2018 Nov 13;6(4):49. doi: 10.3390/jintelligence6040049.

Leutner, F., Codreanu, 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, 36(7), 533-547. doi: 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. doi: 10.1016/j.paid.2016.10.007

Li, Y., & Xie, Y. (2020). Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. Journal of Marketing Research, 57(1), 1–19. doi: 10.1177/0022243719881113

Lignier, B., Petot, J.-M., Canada, B., Pierre De Oliveira, Nicolas, M., Courtois, R., John, O. P., Plaisant, O., & Soto, C. (2022). Factor structure, psychometric properties, and validity of the Big Five Inventory-2 facets: Evidence from the French adaptation (BFI-2-Fr). Current Psychology. doi: 10.1007/s12144-022-03648-0; Q2.

Lippa, R. A. (2010). Gender differences in personality and interests: When, where, and why? Social and Personality Psychology Compass, 4(11), 1098–1110. doi: 10.1111/j.1751-9004.2010.00320.x

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. Psychological Bulletin, 45, 507-529. doi: 10.1037/h0055827.

Loevinger, J., Gleser, C. G., & DuBois, P. H. (1953). Maximising the discriminating power of a multiple score-test. Psychometrika, 18(4), 309-317. doi: 10.1007/BF02289266.

Ludeke, S. G., Bainbridge, T. F., Liu, J., Zhao, K., Smillie, L. D., & Zettler, I. (2019). Using the Big Five Aspect Scales to translate between the HEXACO and Big Five personality models. Journal of Personality, 87(5), 1025–1038. doi: 10.1111/jopy.12453

Maglio, S. J., & Reich, T. (2019). Feeling certain: Gut choice, the true self, and attitude certainty. Emotion, 19(5), 876–888. doi: 10.1037/emo0000490

Malkewitz, C., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. Social sciences & humanities open, 7(1), 100368. doi: 10.1016/j.ssaho.2022.100368

Marcus, B., & Schütz, A. (2005). Who Are the People Reluctant to Participate in Research? Personality Correlates of Four Different Types of Nonresponse as Inferred from Self- and Observer Ratings. Journal of Personality, 73(4), 960–984. doi: 10.1111/j.1467-6494.2005.00335.x

Marsh, H. W., Hau, K., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. Structural Equation Modeling, 11(3), 320-341. doi: 10.1207/s15328007sem1103_2

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. doi: 10.1111/jcal.12170

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 52(1), 81–90. doi: 10.1037/0022-3514.52.1.81

McCrae, R. R., & Costa, P. T., Jr. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (pp. 159–181). The Guilford Press.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis.. British Journal of Mathematical and Statistical Psychology, 23, 1-21. doi: 10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (2013). Test Theory. Psychology Press eBooks. doi: 10.4324/9781410601087

Miles, A., & Sadler-Smith, E. (2014). "With recruitment I always feel I need to listen to my gut": The role of intuition in employee selection. Personnel Review, 43(4), 606–627. doi: 10.1108/PR-04-2013-0065

Momirović, K. (1996). An alternative to Guttman λ6: a measure of true lower bound to reliability of the first principal component. Psihologija, 99-102.

Murphy, M. (2011). Hiring for Attitude: A Revolutionary Approach to Recruiting and Selecting People with Both Tremendous Skills and Superb Attitude. McGraw Hill Professional.

Myszkowski, N., Storme, M., Kubiak, E., & Baron, S. (2022). Exploring the associations between personality and response speed trajectories in low-stakes intelligence tests. Personality and Individual Differences, 191, 111580. doi: 10.1016/j.paid.2022.111580

Myszkowski, N., Storme, M., Kubiak, E., & Baron, S. (in press). The role of personality traits in skipping forced-choice questions: an explanatory item response theory investigation.

Nunnally, J.C. (1978) Psychometric theory. 2nd Edition, McGraw-Hill, New York.

Nunnally, J.C. and Bernstein, I.H. (1994) The Assessment of Reliability. Psychometric Theory, 3, 248-292. doi: 10.12691/education-5-5-2

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. Psychological Methods, 5(3), 343–355. doi: 10.1037/1082-989X.5.3.343

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224–239). The Guilford Press.

Piedmont, R. L., & Hyland, M. E. (1993). Inter-Item Correlation Frequency Distribution Analysis: A Method for Evaluating Scale Dimensionality. Educational and Psychological Measurement, 53(2), 369-378. doi: 10.1177/0013164493053002006

Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. A., & John, O. P. (2010). Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R. Annales médico-psychologiques, 168(2), 97-106. doi: 10.1016/j.amp.2009.09.003

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. A. (2014). Detecting meaning in RSVP at 13 ms per picture. Attention, perception & psychophysics, 76(2), 270-279. doi: 10.3758/s13414-013-0605-z

Raad, B. d., & Perugini, M. (Eds.). (2002). Big five factor assessment: Introduction. In B. de Raad & M. Perugini (Eds.), Big five assessment (pp. 1–18). Hogrefe & Huber Publishers.

Rammstedt, B. (2007). The 10-Item Big Five Inventory. European Journal of Psychological Assessment, 23(3), 193-201. doi: 10.1027/1015-5759.23.3.193

Raykov, T., & Marcoulides, G. A. (2012). Evaluation of validity and reliability for hierarchical scales using latent variable modeling. Structural Equation Modeling, 19(3), 495–508. doi: 10.1080/10705511.2012.687675

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. Multivariate Behavioral Research, 14 (1), 57-74. doi: 10.1207/s15327906mbr1401_4

Revelle, W., & Condon, D. M. (2015). A model for personality at three levels. Journal of Research in Personality, 56, 70–81. doi: 10.1016/j.jrp.2014.12.006

Revelle, W., Zinbarg, R.E. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. Psychometrika 74, 145–154 (2009). doi: 10.1007/s11336-008-9102-z

Roberts, B. W., & Davis, J. P. (2016). Young adulthood is the crucible of personality development. Emerging Adulthood, 4(5), 318–326. doi: 10.1177/2167696816653052

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. Psychological Bulletin, 126(1), 3–25. doi: 10.1037/0033-2909.126.1.3

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. Current Directions in Psychological Science, 17(1), 31–35. doi: 10.1111/j.1467-8721.2008.00543.x

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. Psychological Bulletin, 132(1), 1-25. doi: 10.1037/0033-2909.132.1.1

Rodrigues, R., & Baldi, V. (2017). Interaction mediated by a swipe culture: An observation focused on mobile dating applications. Iberian Conference on Information Systems and Technologies. doi: 10.23919/cisti.2017.7975868

Sackett, P. R., Zhang, C., Berry, C. P. L., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. Journal of Applied Psychology, 107(11), 2040-2068. doi: 10.1037/apl0000994

Schmidt, F., Oh, I.S., & Schaffer, J. (2016). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings. Working Paper.

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

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2009). "Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures": Correction to Schmitt et al. (2008). Journal of Personality and Social Psychology, 96(1), 118. doi: 10.1037/a0014651

Schwaba, T., Rhemtulla, M., Hopwood, C. J., & Bleidorn, W. (2020). A facet atlas: Visualizing networks that describe the blends, cores, and peripheries of personality structure. PLOS ONE, 15(7), e0236893. doi: 10.1371/journal.pone.0236893

Seybert, J., & Becker, D. (2019). Examination of the Test–Retest Reliability of a Forced-Choice Personality Measure. ETS Research Report Series, 2019(1), 1-17. doi: 10.1002/ets2.12273

Shchebetenko, S., Kalugin, A. Y., Mishkevich, A., Soto, C. J., & John, O. P. (2020). Measurement Invariance and Sex and Age Differences of the Big Five Inventory–2: Evidence From the Russian Version. Assessment, 27(3), 472-486. doi: 10.1177/1073191119860901

Short, J. C., McKenny, A. F., & Reid, S. W. (2018). More than words? Computer-aided text analysis in organizational behavior and psychology research. Annual Review of Organizational Psychology and Organizational Behavior, 5(1), 415-435. doi: 10.1146/annurev-orgpsych-032117-104622

Sinclair, S., & Agerström, J. (2020). Does expertise and thinking mode matter for accuracy in judgments of job applicants' cognitive ability ? Scandinavian Journal of Psychology, 61(4), 484-493. doi: 10.1111/sjop.12638

Sijtsma, K. (2009). On the Use, Misuse, and Very Limited Usefulness of Cronbach's Alpha. Psychometrika, 74(1), 107-120. doi: 10.1007/s11336-008-9101-0

Smith, R. W., Min, H., Ng, M. A., Haynes, N. J., & Clark, M. A. (2022). A content validation of work passion: Was the passion ever there? Journal of Business and Psychology, 38(1), 191-213. doi: 10.1007/ s10869-022-09807-1.

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes ? The Life Outcomes of Personality Replication Project. Psychological Science, 30(5), 711-727. doi: 10.1177/0956797619831612

Soto, C. J. (2021). Do Links Between Personality and Life Outcomes Generalize ? Testing the Robustness of Trait–Outcome Associations Across Gender, Age, Ethnicity, and Analytic Approaches. Social Psychological and Personality Science, 12(1), 118-130. doi: 10.1177/1948550619900572

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology, 100(2), 330-348. doi: 10.1037/a0021717

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. Journal of Personality and Social Psychology, 113(1), 117-143. doi: 10.1037/pspp0000096.

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. Journal of Research in Personality, 68, 69-81. doi: 10.1016/j.jrp.2017.02.004

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity ? Psychological Assessment, 31(4), 444-459. doi: 10.1037/pas0000586

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. The American Journal of Psychology, 15(2), 201–293. doi: 10.2307/1412107

Steiger, J. H., & Lind, J. C. (1980). Statistically-Based Tests for the Number of Common Factors. doi: 10.12691/rpbs-4-1-3

Tang, W. et Cui, Y. (2012). A simulation study for comparing three lower bounds to reliability. Communication présentée à l'AERA Division D: Measurement and research methodology, section 1: Educational measurement, psychometrics, and assessment (1-25).

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the multidimensional personality questionnaire. In The SAGE handbook of personality theory and assessment, vol 2: Personality measurement and testing (pp. 261-292). Sage Publications, Inc. doi: 10.4135/ 9781849200479.n13

Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. Annual Review of Organizational Psychology and Organizational Behavior, 8, 199–233. doi: 10.1146/annurev-orgpsych-012420-062228

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. Psychological Bulletin, 146(1), 30-90. doi: 10.1037/bul0000217

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the Maximal Split-Half Coefficient to Estimate Reliability. Educational and Psychological Measurement, 70(2), 232-251. doi: 10.1177/0013164409355688

Thorndike, E. L. (1918). Individual differences. Psychological Bulletin, 15(5), 148–159. doi: 10.1037/h0070314

Thurstone, L.L. (1947). Multiple factor analysis. University of Chicago Press: Chicago.

Trizano-Hermosilla, I. et Alvarado, J. M. (2016). Best alternatives to Cronbach's Alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. Frontiers in psychology, 7(769), 1-8. doi: 10.3389/fpsyg.2016.00769

Valentino, N. A., Zhirkov, K., Hillygus, D. S., & Guay, B. (2021). The Consequences of Personality Biases in Online Panels for Measuring Public Opinion. Public Opinion Quarterly, 84(2), 446-468. doi: 10.1093/poq/nfaa026

Van Der Ark, L. A., Van Der Palm, D. W., & Sijtsma, K. (2011). A Latent Class Approach to Estimating Test-Score Reliability. Applied Psychological Measurement, 35(5), 380-392. doi: 0.1177/0146621610392911

Vedel, A., Wellnitz, K. B., Ludeke, S., Soto, C. J., John, O. P., & Andersen, S. C. (2021). Development and validation of the Danish Big Five Inventory-2: Domain- and facet-level structure, construct validity, and reliability. European Journal of Psychological Assessment, 37(1), 42–51. doi: 10.1027/1015-5759/a000570

von Davier, M. (2018). Automated item generation with recurrent neural networks. Psychometrika, 83(4), 847- 857. doi: 10.1007/s11336-018-9608-y

Walker, D. A. (1931). Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 22, 73–86.

Weidner, N.W. and Landers, R.N. (2020), "Swipe right on personality: a mobile response latency measure", Journal of Managerial Psychology, Vol. 35 No. 4, pp. 209-223. doi: 10.1108/JMP-07-2018-0330

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender Differences in Personality across the Ten Aspects of the Big Five. Frontiers in Psychology, 2. doi: 10.3389/fpsyg.2011.00178

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), The ITC international handbook of testing and assessment (pp. 349–363). Oxford University Press. doi: 10.1093/med:psych/9780199356942.003.0024

Will, P., Krpan, D. & Lordan, G. People versus machines: introducing the HIRE framework. Artif Intell Rev 56, 1071–1100 (2023). doi: 10.1007/s10462-022-10193-6

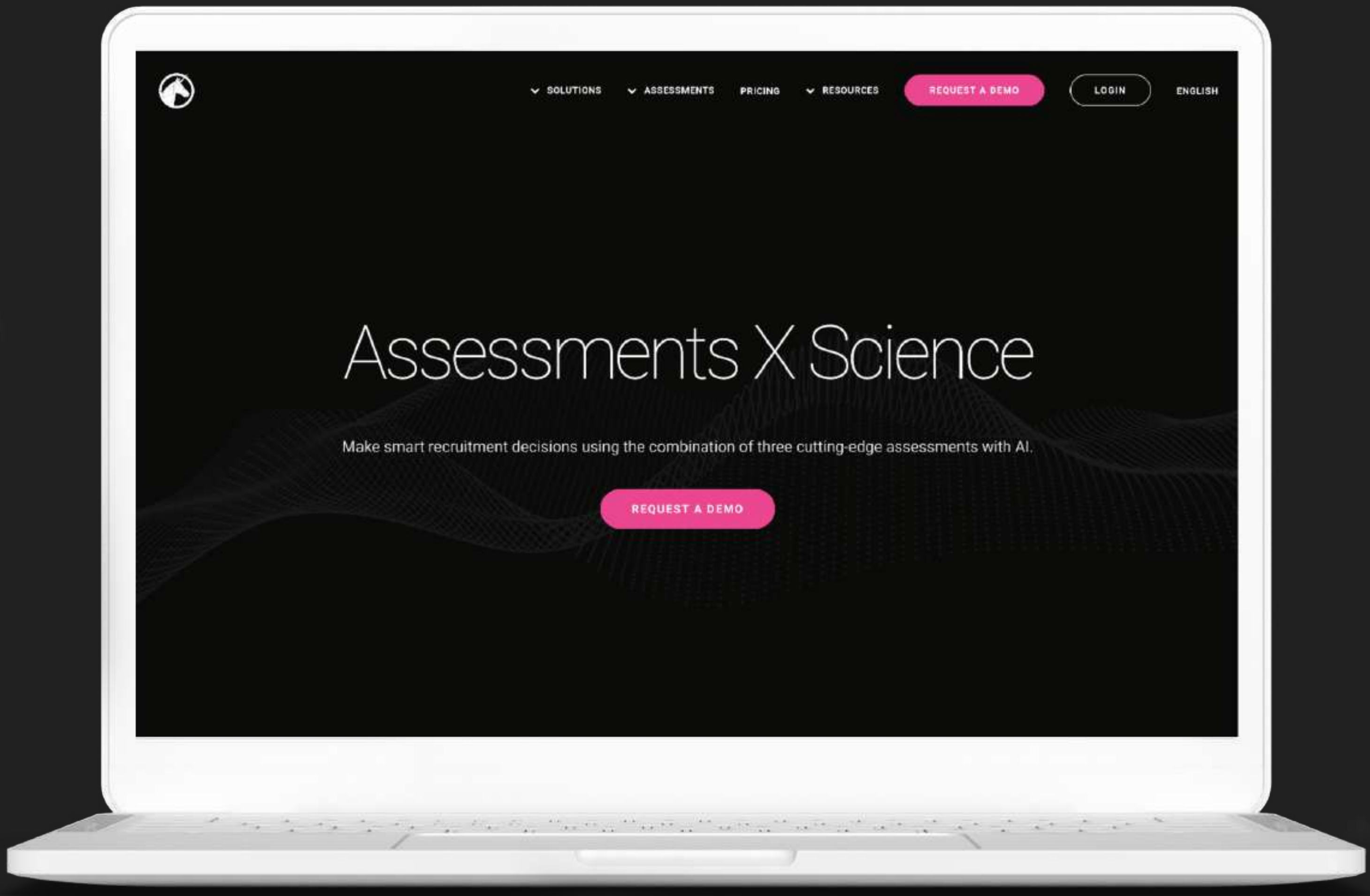
Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806-838. doi: 10.1177/ 0011000006288127

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. American Psychologist, 70(1), 10-20. doi: 10.1037/a0038208

Zhang, B., Li, Y., Li, J., Luo, J., Ye, Y., Yin, L., Chen, Z., Soto, C. J., & John, O. P. (2021). The Big Five Inventory–2 in China: A Comprehensive Psychometric Evaluation in Four Diverse Samples. Assessment, 29(6), 1262-1284. doi: 10.1177/10731911211008245

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's, Revelle's, and McDonald's: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70, 123- 133. doi: 10.1007/s11336-003-0974-7

Zinko, R., Stolk, P., Furner, Z., & Almond, B. (2020). A picture is worth a thousand words: how images influence information quality and information load in online reviews. Electronic Markets, 30(4), 775-789. doi: 10.1007/s12525-019-00345-y



ASSESSFIRST

AssessFirst has developed a predictive recruitment solution that allows companies to predict the extent to which candidates and employees will succeed and thrive at work. The AssessFirst solution analyses data from over 5,000,000 profiles. More than 3,500 companies use the solution to increase their performance by up to 40%, reduce their recruitment costs by 20%, and cut their employee turnover rates by 50%. This signifies a substantial impact on operational efficiency and cost-effectiveness for organisations that incorporate AssessFirst's tools into their recruitment and human resources management processes.