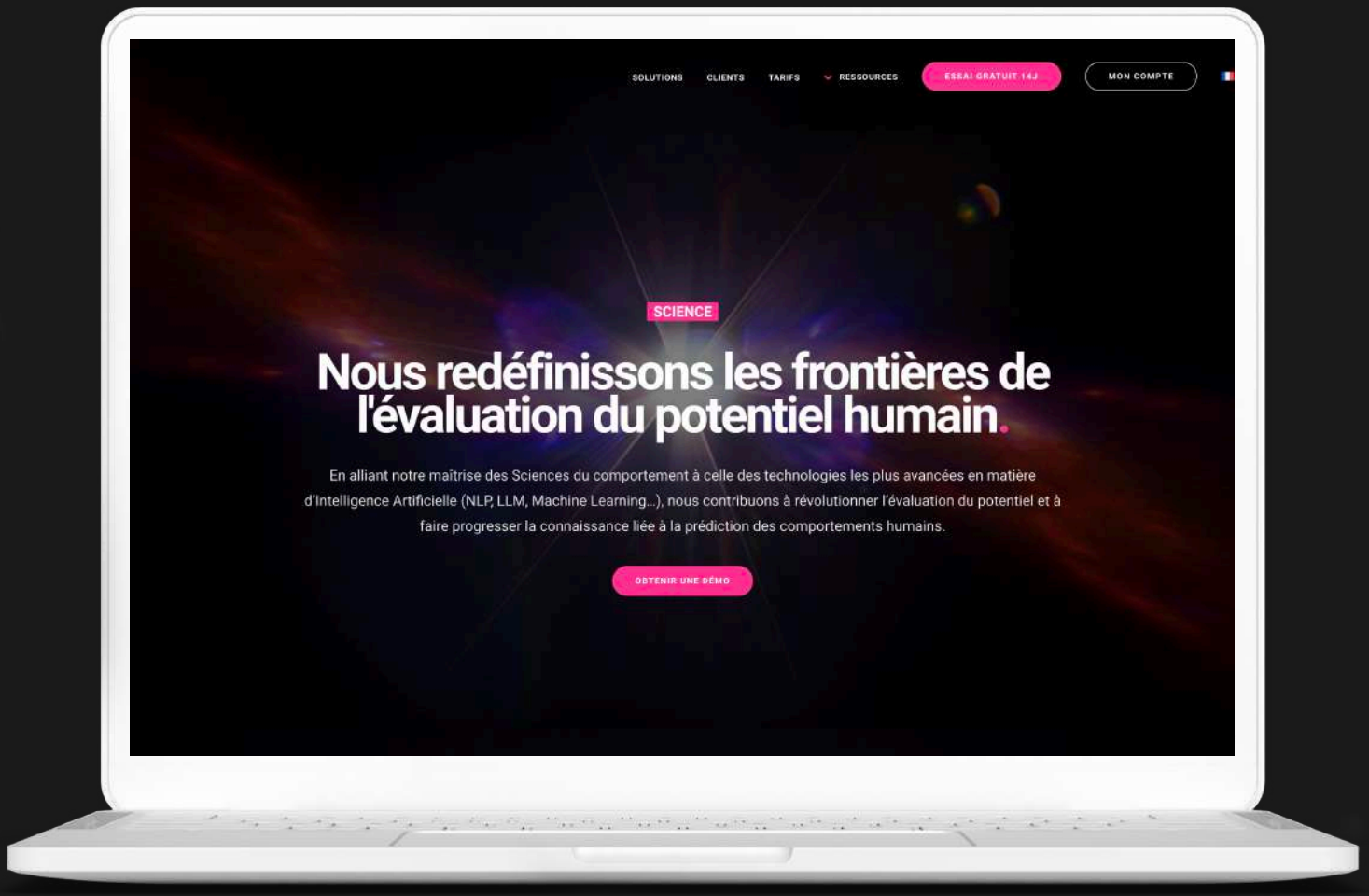


L'évaluation du potentiel par AssessFirst

Ce document fait la synthèse des études psychométriques liées à nos questionnaires de personnalité SWIPE, de motivations DRIVE et de raisonnement BRAIN. Pour chacun de ces questionnaires, sont présentées des informations liées à leur format et à leur construction, ainsi qu'à des notions psychométriques clés : validité, fidélité, sensibilité et équité des résultats aux questionnaires.

ASSESSFIRST X SCIENCE

Préface

Prendre de bonnes décisions RH ne s'improvise pas. Si de nombreuses entreprises font évoluer leurs pratiques pour le meilleur, certaines continuent de s'appuyer sur des méthodologies peu fiables pour sélectionner les candidats. Dans l'histoire du recrutement, l'entretien non structuré a en ce sens été la technique la plus utilisée (Buckley, Norris & Wiese, 2000), car perçu comme plus efficace, professionnelle et agréable que les autres outils (Highhouse, 2008). Cette façon de procéder a toutefois contribué à détériorer la qualité des décisions, en laissant une trop grande place à l'intuition, aux préjugés et aux biais cognitifs (Sinclair & Agerström, 2020; Miles & Sadler-Smith, 2014; Ames, Kammrath, Suppes & Bolger, 2010). En témoignent, par exemple, les nombreux échecs de recrutement (Murphy, 2011) ou encore la persistance de nombreuses discriminations dans les processus de sélection (Benson, Li & Shue, 2022; Kessler, Low & Sullivan, 2019). C'est pourquoi il est important de prendre le temps de mesurer les critères qui donneront une prédiction solide de la capacité d'un candidat à réussir, plutôt que de céder à la simplicité et la satisfaction immédiate qu'amène une prise de décision intuitive (Maglio & Reich, 2019; Kirkebøen & Nordbye, 2017). Les recherches en psychologie du travail démontrent ainsi que (1) la personnalité, les motivations et les capacités de raisonnement s'avèrent de meilleurs facteurs prédictifs de la performance en poste (Sackett, Zhang, Berry & Lievens, 2023; Sackett, Zhang, Berry & Lievens, 2021; Schmidt, Oh & Shaffer, 2016), (2) une simple équation s'avère plus performante et précise pour recruter efficacement (Will, Krpan & Lordan, 2022; Kuncel, Klieger, Connelly & Ones, 2013), ou encore que (3) les entreprises qui suivent les recommandations des tests - notamment de personnalité et de raisonnement, réalisent de meilleurs recrutements (Hoffman, Kahn & Li, 2015).

C'est dans ce cadre qu'AssessFirst développe et distribue des outils de mesure psychométrique avec un objectif en tête : permettre aux professionnels des ressources humaines de disposer d'indicateurs fiables et objectifs relatifs aux caractéristiques psychologiques et comportementales des personnes en contexte professionnel. AssessFirst lie à la fois des outils issus de la psychologie comportementale (qui ont été développés et validés par des équipes de psychologues du travail et de psychométriciens selon les standards internationaux les plus exigeants) et une technologie IA. La mise en conformité de ces outils avec les standards préconisés par l'Association Américaine de Psychologie (A.P.A.) et la Commission Internationale des Tests (I.T.C.) permet aujourd'hui à AssessFirst de garantir un haut niveau de qualité dans la conception des questionnaires et d'améliorer continuellement la fidélité de ses outils d'évaluation.

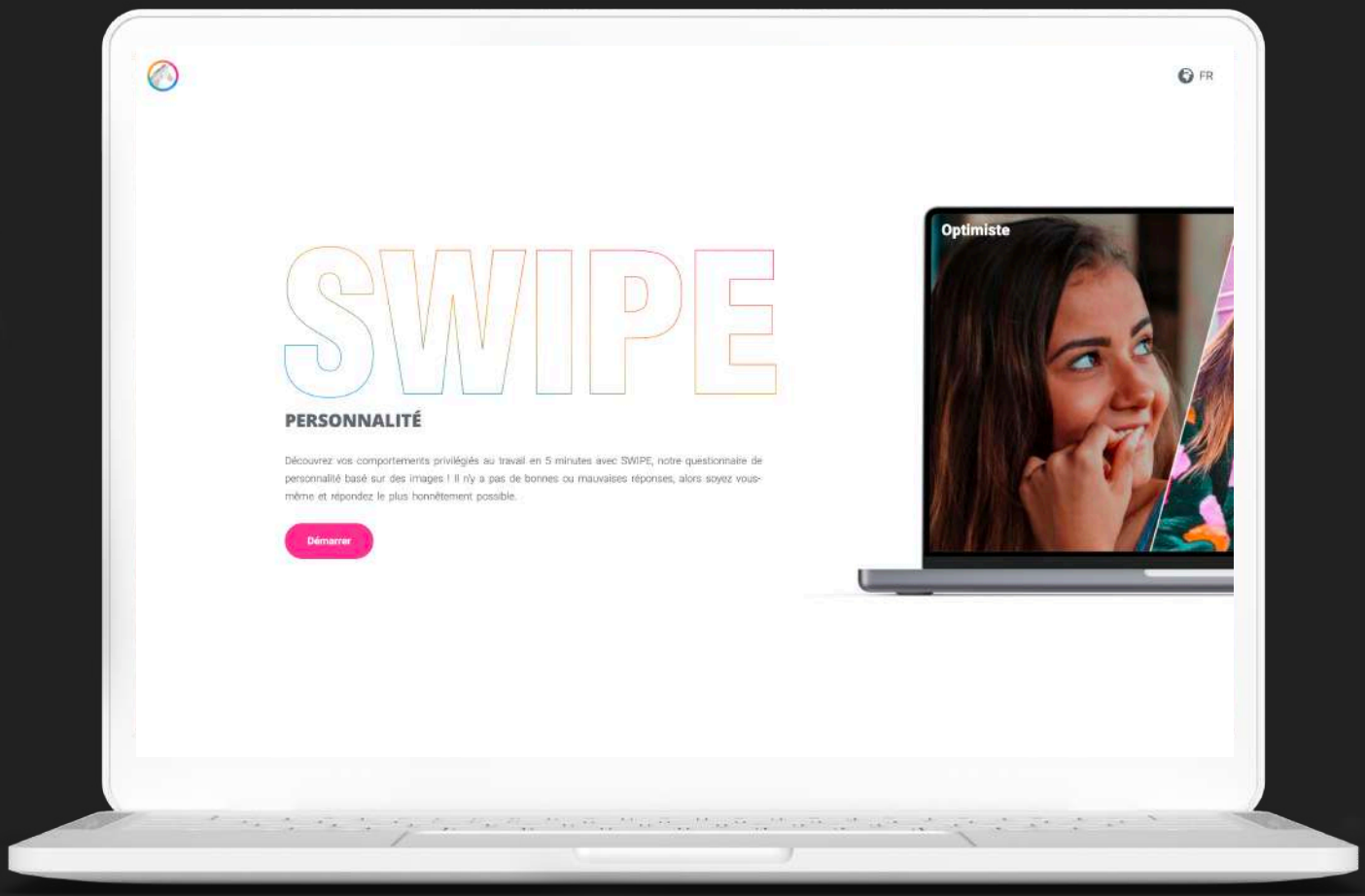
Ce document fait la synthèse des études psychométriques liées à nos questionnaires de personnalité SWIPE et SHAPE, de motivations DRIVE et de raisonnement BRAIN. Pour chacun de ces questionnaires, sont présentées des informations liées à leur format et à leur construction, ainsi qu'à des notions psychométriques clés : validité, fidélité, sensibilité et équité des résultats aux questionnaires. En ce sens, il permet de justifier de la qualité et de la pertinence des outils d'évaluation proposés par AssessFirst.

Sommaire

SWIPE

DRIVE

BRAIN



Bienvenue dans le futur de l'évaluation de la personnalité

SWIPE est un court questionnaire de personnalité, basé sur des images, qui vous aide à mieux comprendre la façon unique dont une personne se comporte en situation professionnelle. Pensé mobile-first et réalisable en 5 minutes, SWIPE allie le meilleur de la psychométrie moderne et de l'expérience utilisateur.

SWIPE

1. Introduction

SWIPE est un court questionnaire de personnalité, basé sur des images, qui vous aide à mieux comprendre la façon unique dont une personne se comporte en situation professionnelle. Pensé mobile-first et réalisable en **5 minutes,** SWIPE allie le meilleur de la psychométrie moderne et de l'expérience utilisateur. Ce questionnaire est composé de **72 items** mesurant 6 traits et **18 facettes de personnalité**, et de 3 items de recueil de données, pour un total de 75 items. SWIPE a été conçu en 2023 par l'équipe Science d'AssessFirst et, à date, a déjà fait l'objet de **6 publications** dans des congrès internationaux de psychologie ou revues scientifiques à comité de lecture.

2. Historique de développement

Alors que la personnalité est un critère déterminant dans la prédiction de la réussite en poste (Judge & Zapata, 2015), les questionnaires de personnalité actuellement proposés sur le marché sont, dans de nombreux cas, considérés comme longs, peu modernes et ne permettant pas de favoriser une bonne expérience utilisateur. En ce sens, les questionnaires traditionnels reçoivent généralement des scores moyens de favorabilité (Hausknecht, Day & Thomas, 2004). Ces conclusions semblent légitimes dans un monde où tout est plus rapide, visuel et mobile-first : Instagram pour partager des photos, Spotify pour écouter une musique, Google Maps pour trouver son chemin en quelques clics, ou encore Tinder pour trouver l'amour en swipant[1]. Ces codes et moyens d'accéder à l'information sont aujourd'hui acceptés par tous, et constituent nos modes d'interactions avec le numérique et ses capacités. Aussi, pour répondre à la volonté des candidats de se voir proposer des processus de recrutement qui soient plus rapides et entièrement réalisables via smartphone (Böhm & Jäger, 2016), les entreprises se doivent d'intégrer ces nouvelles normes pour rester attractives et compétitives. C'est en marge de ces constats, et face aux besoins toujours plus importants des professionnels RH de concevoir des processus de décision qui soient rapides, fiables et engageants, que SWIPE a été pensé et développé (Kubiak, Niesner & Baron, 2023). Le développement de SWIPE a ainsi été animé par 4 besoins ou postulats.

2.1. Rapide

La durée idéale d'un questionnaire de personnalité est un sujet complexe, qui doit assurer le respect des attentes des candidats, la perception d'équité et de sérieux de la mesure, et la validité de l'évaluation. Aussi, l'objectif n'est pas forcément de proposer un questionnaire qui soit le plus rapide possible, mais plutôt de trouver le point d'équilibre parfait permettant d'optimiser notre réponse à ces besoins. À ce titre, les études scientifiques sur le sujet concluent que, (1) les candidats préfèrent un temps d'assessment global compris entre 10 et 30 minutes, et qu'il n'existe pas de déperdition au-delà de 20 minutes (Hardy, Gibson, Sloan & Carr, 2017), (2) que les questionnaires trop longs peuvent perdre en validité (Burisch, 1997) et la mesure être influencée par d'autres facteurs (Myszkowski, Storme, Kubiak & Baron, 2022), (3) que les questionnaires trop courts, bien qu'utiles, ne permettent pas de capturer toute l'information sur la personnalité d'une personne (Hofmans, Kuppens & Allik, 2008), (4) que le nombre idéal de points de mesure par échelle se situe entre 6 et 9 items ou proposals (Soto & John, 2019). Pour le développement de SWIPE, nous avons ainsi maximisé ces éléments, en construisant un questionnaire comprenant 8 proposals principales par facette, et d'une durée moyenne de 5 minutes, afin d'atteindre un temps global d'évaluation d'environ 25 minutes, prenant également en compte nos questionnaires DRIVE et BRAIN.

[1] Swiper : balayage de l'écran.

2.2. Mobile first

La manière dont les candidats souhaitent compléter un questionnaire d'évaluation dans le cadre d'un processus de recrutement a considérablement changé au cours des dernières années, et s'oriente davantage vers l'usage du mobile (Lawrence & Kinney, 2017; Smith, 2015). Ce mode d'administration présente plusieurs avantages en permettant notamment : (1) de répondre aux évolutions sociétales et technologiques, qui font aujourd'hui du mobile le principal outil de consommation média (Goovaerts, 2016) et d'accès à internet (Smith, 2015), (2) de proposer la complétion des questionnaires n'importe où et n'importe quand (Arthur & Traylor, 2019), (3) d'ouvrir davantage la sélection aux candidats issus de groupes historiquement discriminés, au sens que les évaluations via mobile sont plus facilement complétées par les femmes ou les populations afro-américaines et hispaniques (Arthur, Doverspike, Muñoz, Taylor & Carr, 2014). Néanmoins, il serait naïf de simplement chercher à convertir un questionnaire de personnalité « computer-based » pour le proposer, en l'état, sur mobile. En ce sens, déployer une évaluation sur mobile sans s'assurer de sa pleine adaptation et accessibilité contribue à détériorer l'expérience utilisateur (Gutierrez & Meyer, 2013). En revanche, quand ces évaluations sont pensées mobile-first et initialement conçues pour un déploiement en environnement mobile, les réactions des candidats sont identiques quel que soit le support utilisé (Kinney, Lawrence et Chang, 2014). En conséquence, SWIPE a été pensé mobile-first grâce à nos équipes de psychologues et de UX designer.

2.3. Engageant

Si la rapidité de SWIPE et sa conception mobile-first contribuent à le rendre plus engageant et apprécié des utilisateurs, trois éléments principaux sont au cœur de la qualité de l'expérience de SWIPE, à savoir :

- **La gamification par l'image** : alors que la personnalité est souvent mesurée grâce à des questionnaires utilisant des échelles de Likert classiques, les nouvelles tendances et possibilités technologiques font de la gamification un levier d'engagement des utilisateurs (Leutner, Akhtar & Chamorro-Permuzic, 2022; Leutner & Chamorro-Permuzic, 2018; Armstrong, Ferrell, Collmus & Landers, 2016; Chamorro-Permuzic, Winsborough, Sherman & Hogan, 2016). En ce sens, les évaluations gamifiées sont perçues comme étant plus immersives que les évaluations traditionnelles (Leutner, Codreanu, Liff & Mondragon, 2020), permettent de diminuer l'anxiété des utilisateurs (Mavridis & Tsiatsos, 2016) et d'augmenter leur satisfaction : en résulte une plus forte perception d'équité du processus du recrutement et une meilleure attractivité organisationnelle lorsque ces évaluations sont utilisées (Georgiou & Nikolaou, 2020). Parmi les différents moyens de gamification, l'utilisation d'images pour mesurer la personnalité se révèle une stratégie efficace (Hilliard, Kazim, Bitsakis & Leutner, 2022; Leutner, Codreanu, Liff & Mondragon, 2020; Krainikovsky, Melnikov & Samarev, 2019; Leutner, Yearsley, Codreanu, Borenstein & Ahmetoglu, 2017), au sens qu'elle permet à la fois une mesure valide de la personnalité (Kubiak, Niesner & Baron, 2023) et d'optimiser la satisfaction des utilisateurs (Efremova, Kubiak & Baron, 2023). En effet, l'utilisation d'images permet d'apporter davantage de contexte à l'item et d'information à l'utilisateur, de faciliter et d'accélérer sa lecture - les images étant traitées beaucoup plus rapidement que le texte par le cerveau humain (Potter, Wyble, Hagmann & McCourt, 2014), ou encore de collecter d'autres points de données qui permettent d'inférer des indices quant à la personnalité du répondant (Kubiak, Bernard & Baron, 2023). Toutefois, plutôt qu'uniquement proposer des images, SWIPE s'inspire aussi des recherches en marketing, consommation et sciences de la décision, qui montrent que les formats hybrides - combinant un court texte et une image - sont plus performants (Wu, Wu & Wang, 2020), que les images permettent d'apporter plus d'informations et de dépasser les biais de textes non-lus car trop longs (Zinko, Stolk, Furner & Almond, 2019), ou encore que les images de qualité professionnelle, qui représentent un humain, et où l'association texte-image est optimale, permettent de générer l'engagement des utilisateurs sur certains médias sociaux (Lie & Xie, 2019). Capitalisant sur ces conclusions, SWIPE est ainsi un questionnaire « image-based », mais qui associe un court texte descriptif à chaque image, afin de maximiser la qualité de l'information et l'engagement des répondants.

- **Le « swipe » comme moyen de réponse** : l'émergence de la consommation via mobile a aussi contribué à intégrer de nouveaux moyens d'interaction physique avec cette information. Parmi ceux-ci, le « swipe » - touché de l'écran suivi d'un mouvement de glissement, s'est imposé comme l'un des mouvements les plus utilisés par les concepteurs d'applications mobiles, et fait partie de notre quotidien. Au sens que le « swipe » permet de simplifier les actions et décisions sur mobile en les rendant binaires, de faire les choses plus rapidement (Rodrigues & Baldi, 2017), ou encore qu'il s'avère plus fluide, intuitif et compréhensible pour les utilisateurs, il contribue à accroître leur satisfaction (Dou & Sundar, 2016). En somme, la logique du swipe, si elle est avant tout technique et physique, se pose aussi en levier de satisfaction et de persuasion psychologique (David & Cambre, 2016). Dans le domaine de l'évaluation de la personnalité, de nouvelles recherches ont commencé à mettre en avant les effets bénéfiques du swipe pour l'engagement des utilisateurs (Efremova, Kubiak & Baron, 2023) et la réduction du temps de réponse par item (Weidner & Landers, 2020). Si les utilisateurs peuvent répondre au questionnaire SWIPE grâce à différents moyens - notamment dans la version desktop, le mouvement de swipe est ainsi le moyen unique de complétion du questionnaire dans sa version mobile.

- **Le format de réponse** : les formats de réponse à choix forcé - le répondant doit choisir entre 2 options - gagnent en popularité et se positionnent comme une alternative aux mesures de type Likert ou à déclaration unique. Notamment, le choix forcé permet de neutraliser les biais d'acquiescement, d'extrémité ou d'auto-complaisance - sévérité (Wetzel, Böhnke & Brown, 2016; Paulhus & Vazire, 2007), ou encore de drastiquement réduire les tentatives de triches (Cao & Drasgow, 2019). Couplés à des modèles de scoring IRT (Brown & Maydeu-Olivares, 2011), les formats de réponse à choix forcé s'avèrent donc plus efficaces pour mesurer la personnalité. Néanmoins, ce format de réponse peut être cognitivement plus lourd pour les utilisateurs, conduisant à une prise de décision plus compliquée et à une expérience critiquée (Bartram & Brown, 2004). Bénéficier des avantages du choix forcé, tout en améliorant l'expérience des utilisateurs grâce à ce format, appelle ainsi à des actions adaptatives pour dépasser les réactions mitigées naturellement inhérentes à ce format. Aussi, trois types de correctifs, qui ont démontré leurs effets pour améliorer l'engagement des utilisateurs face à ce format de réponse (Dalal, Zhu, Rangel, Boyce & Lobene, 2021), ont été pris en compte pour développer SWIPE :

  - (1) Apporter plus de latitude à l'aspect dichotomique de ce format : la critique des formats à choix forcé réside dans le fait que les utilisateurs souhaiteraient sélectionner les deux propositions, ou aucune. Aussi, l'absence de cette possibilité dans les questionnaires actuels génère de la frustration (Bartram & Brown, 2004). Pour y remédier, SWIPE donnera aux utilisateurs la possibilité, un certain nombre de fois, de sélectionner les deux options de réponse, ou aucune. Nos études montrent que ce double-choix est bénéfique pour l'expérience utilisateur (Efremova, Kubiak & Baron, 2023) et apporte de l'information sur la personnalité du répondant (Baron, Storme, Myszkowski & Kubiak, 2023; Myszkowski, Storme, Kubiak & Baron, in press) ;

  - (2) Inclure un feedback suite à la passation : une synthèse suite à la passation de SWIPE est proposée au répondant. Celle-ci est générée automatiquement et apporte des éléments concrets de compréhension au répondant quant à sa personnalité, ses comportements privilégiés, son style personnel et ses axes d'amélioration. L'ensemble de ces contenus est tourné positivement et a pour objectif de permettre au répondant de mieux se connaître, de manière simple et objective : aussi, 90% des utilisateurs trouvent cette synthèse facile à comprendre, et 98% la trouve utile[2] ;

  - (3) Retirer les items les moins socialement désirables : si un équilibre entre items formulés positivement et négativement est nécessaire pour assurer une bonne validité au questionnaire (Soto & John, 2019), nous avons veillé à retirer les items et propositions de réponse qui étaient jugés beaucoup trop négativement, et ainsi jamais sélectionnées par les répondants. Cette stratégie permet de ne pas inclure d'items où le répondant aurait à faire un choix complexe ou psychologiquement embarrassant.

[2] Enquête qualitative réalisée auprès de 180 utilisateurs.

2.4. Fiable

Parmi l'ensemble des modèles de personnalité existant, le modèle des Big Five a depuis de nombreuses années, et à travers plusieurs milliers d'études, démontré sa validité, sa fidélité et son utilité (Goldberg, 1993b; John, Naumann, & Soto, 2008; McCrae & Costa, 2008). Notamment, les traits de personnalité du Big Five ont de manière constante permis de prédire la performance au travail (Barrick & Mount, 1991; Judge, Higgins, Thoresen & Barrick, 1999; Higgins, Peterson, Pihl & Lee, 2007; Kuncel, Ones & Sackett, 2010; Schmitt, 2014), et encore plus quand celle-ci est contextualisée aux exigences d'un métier spécifique (Judge & Zapata, 2015; Tett, Toich & Ozkum, 2021). Aussi, la popularité de ce modèle amène la communauté scientifique a constamment le challenger, et à l'améliorer de manière continue : les recherches sur le « Big Five Inventory-2 » ont ainsi permis d'introduire une structure hiérarchique plus robuste au modèle, et d'améliorer sa fidélité et son pouvoir prédictif, tout en conservant l'orientation conceptuelle et la facilité de compréhension du modèle d'origine (Soto & John, 2017a). De même, encore plus récemment, le modèle BFI-2 a évolué : ce modèle étant historiquement considéré comme sous-optimal pour évaluer l'échelle Honesty-Humility (H) de l'HEXACO, de nouvelles recherches ont ainsi proposé l'ajout de 3 facettes ad hoc pour la mesure de cette échelle H (Denissen, Soto, Geenen, John & Van Aken, 2022; Lee, Ashton & De Vries, 2022). Afin d'assurer le développement d'un outil qui repose sur les modèles de personnalité les plus efficaces et modernes, SWIPE a donc été pensé pour maximiser la validité convergente avec le BFI-2 + sa nouvelle échelle d'humilité. Plus de détails dans la section suivante.

3. Bases théoriques

3.1. Le modèle des Big Five et son évolution

Le questionnaire de personnalité SWIPE repose sur le modèle des Big Five (Goldberg, 1993b; John, Naumann, & Soto, 2008; McCrae & Costa, 2008). Ce modèle, aussi nommé modèle des cinq facteurs (FFM : Five Factor Model), a initialement été développé suite à une analyse factorielle d'un grand nombre de rapports d'évaluation relatifs à des adjectifs et items de questionnaires de personnalité : d'un point de vue lexical, le développement du FFM (Digman, 1990; Goldberg, 1992; John, 1990; McCrae & Costa, 1987) est basé sur plusieurs décennies de recherches. Le modèle des Big Five identifie ainsi cinq grands traits de personnalité : l'Extraversion, l'Agréabilité, l'Ouverture à l'expérience - ou Ouverture, la Conscienciosité, et le Stabilité émotionnelle - aussi appelée Névrosisme. Voir les détails dans le tableau 3.1.

Traits	Définitions	ACL marker items
Extraversion	Degré auquel la personne a besoin d'interactions sociales avec les autres.	Quiet, Reserved, Shy vs. Talkative, Assertive, Active
Agréabilité	Degré auquel la personne développe des relations harmonieuses avec les autres.	Fault-finding, Cold, Unfriendly vs. Sympathetic, Kind, Friendly
Ouverture	Degré auquel la personne se challenge intellectuellement et se montre curieuse.	Commonplace, Narrow-interest, Simple vs. Wide-interest, Imaginative, Intelligent
Conscienciosité	Degré auquel la personne a tendance à se conformer à des normes et standards.	Careless, Disorderly, Frivolous vs. Organized, Through, Precise
Stabilité émotionnelle	Degré auquel la personne a tendance à ressentir des émotions négatives.	Tense, Anxious, Nervous vs. Stable, Calm, Contented

Table 3.1. Traits des Big Five et leur description.

Ce modèle est le meilleur point de départ pour le développement d'outils de mesure de la personnalité. En effet, les études montrent que tous les inventaires multidimensionnels de personnalité peuvent être reconfigurés autour de ces cinq grands traits (Raad & Perugini, 2002). Autrement dit, l'ensemble des différences inter-individuelles quant aux comportements, sentiments et modes de pensées peuvent à ce jour être résumées à ces cinq traits. Depuis son émergence, ce modèle a réussi à acquérir une forte solidité et reconnaissance scientifique : aussi, plusieurs dizaines d'années de recherches ont contribué à son amélioration, et à la création de questionnaires de mesures validés, comme le « Big Five Inventory » - aussi nommé « BFI », composé de 44 items de type Likert (John, Donahue, & Kentle, 1991).

Néanmoins, les 30 années écoulées depuis la création du BFI ont permis d'affiner et de préciser notre compréhension de la personnalité, de sa structure et de son évaluation. Aussi, des travaux plus récents ont permis d'intégrer cette connaissance nouvelle, tout en remédiant aux limites structurelles et psychométriques identifiées à la première version du BFI. Ces recherches ont ainsi permis d'aboutir à la publication du « Big Five Inventory - 2 » ou « BFI-2 » (Soto & John, 2017a), et permettent d'introduire une structure hiérarchique plus robuste au modèle - composé de 15 facettes de personnalité, d'améliorer sa fidélité et son pouvoir prédictif, tout en conservant l'orientation conceptuelle et la facilité de compréhension du modèle d'origine. Ces résultats font du « BFI-2 » une mesure fiable et valide des traits du Big Five et de leurs facettes associées, et indiquent que cette seconde version de l'inventaire représente une avancée importante par rapport à la version BFI originale. La version finale du « BFI-2 » est ainsi composée de 60 items de type Likert. D'autres versions ont également été développées : le « BFI-2-S » composé de 30 items, et le « BFI-2-XS » composé de 15 items (Soto & John, 2017b). L'inventaire a aussi fait l'objet récent d'une adaptation française, qui a été utilisé dans le cadre du développement de SWIPE (Lignier, Petot, Canada, Oliveira, Nicolas, Courtois, John, Plaisant & Soto, 2022).

Certaines études suggèrent toutefois que les échelles du « BFI » et du « BFI-2 » sont relativement faibles pour capturer la variance liée à l'humilité, qui s'avère un trait de personnalité important, notamment en contexte professionnel. Aussi, les corrélations obtenues par les Big Five pour prédire la sphère « Honesty-Humility » de l'HEXACO-PI-R semblent particulièrement faibles pour le « BFI-2 » par rapport à celles d'autres inventaires comme le NEO-PI-R (Ashton, Lee & Visser, 2019; Ashton & Lee, 2019; Lee & Ashton, 2019). Ces conclusions sont partagées par des analyses plus récentes, qui mettent en avant la pauvreté du « BFI-2 » pour rendre compte de cette échelle « H » liée à l'humilité. Considérant l'impact de cette échelle dans la prédiction des comportements pro-sociaux (Thielmann, Spadaro & Balliet, 2020), il apparaît ainsi nécessaire de prolonger le modèle des Big Five et du « BFI-2 », en y intégrant un trait lié à l'humilité. De nouvelles recherches ont ainsi proposé l'ajout de 3 facettes ad hoc pour la mesure de cette échelle H au modèle du « BFI-2 » (Denissen, Soto, Geenen, John & Van Aken, 2022; Lee, Ashton & De Vries, 2022).

3.2. Les facettes de personnalité de SWIPE

Les Big Five ont de manière consistante démontré leur intérêt pour prédire la réussite en poste, ou de nombreux éléments de la vie de tous les jours (Soto, 2019; Soto, 2021). SWIPE a donc été développé avec pour objectif de maximiser la validité convergente avec le « BFI-2 » et son échelle ajoutée d'humilité. Afin de permettre d'apporter une lecture fine de chaque profil, une approche par facette est privilégiée : simplement, une facette de personnalité est un schéma caractéristique de pensée, de sentiment ou de comportement qui a tendance à être stable dans le temps et dans les situations (Allport, 1961; Bleidorn, Schwaba, Zheng, Hopwood, Sosa, Roberts & Briley, 2022). Au final, ce sont ainsi 6 traits et 18 facettes de personnalité qui sont mesurés par SWIPE. Ceux-ci sont présentés et définis dans le tableau 3.2.

Traits	Facettes	Définition
EXTRAVERSION	Assertiveness	Tendance à se comporter en leader et à influencer les autres
	Energy level	Tendance à montrer de l'enthousiasme et de l'énergie
	Sociability	Tendance à aller facilement vers les autres, à se montrer sociable et extraverti
AGRÉABILITÉ	Compassion	Tendance à se montrer bienveillant et compatissant avec les autres
	Respectfulness	Tendance à se montrer respectueux, poli et à éviter les conflits
	Trust	Tendance à faire facilement confiance aux autres et à les pardonner
HUMILITÉ	Greed avoidance	Tendance à s'attacher aux choses simples, à être peu matérialiste
	Modesty	Tendance à faire preuve de modestie et d'humilité
	Sincerity	Tendance à faire preuve de sincérité et à se montrer honnête
OUVERTURE	Aesthetic sensitivity	Tendance à s'intéresser à l'art sous toutes ses formes
	Creative imagination	Tendance à faire preuve d'inventivité, à se montrer créatif et original
	Intellectual curiosity	Tendance à faire preuve de curiosité et à s'intéresser aux choses abstraites
CONSCIENCIOSITÉ	Organization	Tendance à s'organiser avec méthode et à faire preuve de méthode
	Productiveness	Tendance à chercher la performance maximale et à être efficace
	Responsibility	Tendance à se montrer fiable et à respecter ses engagements
STABILITÉ ÉMOTIONNELLE	Anxiety	Tendance à ressentir le stress et à se montrer réactif
	Depression	Tendance à ressentir des émotions principalement négatives
	Emotional volatility	Tendance à exprimer et partager ses émotions et ses ressentis.

Table 3.2. Facettes de personnalité mesurés par SWIPE.

4. Construction de SWIPE

Le développement de SWIPE s'est concrétisé à travers un process en 3 phases : (1) la création et le test de plusieurs séries d'items, (2) la sélection des meilleurs items testés, (3) la création et le test d'une série de validation composée des items sélectionnés. Ces 3 étapes sont ci-après présentées plus en détails.

4.1. Phase 1 : séries de test

Entre Mai et Décembre 2022, 6 séries d'items de tests ont été créées et diffusées. Ces séries avaient pour objectif de collecter de la donnée sur un maximum d'items SWIPE « censés » mesurer des facettes de personnalité du BFI-2 : censé, car, bien entendu, nombre de ces items testés ne se sont pas révélés mesurer efficacement la facette qu'ils étaient censés mesurer. Les séries ont été lancées en séquence : quand suffisamment de données avaient été collectées sur la série 1, la série 2 était mise en production, etc. Afin de recueillir de la donnée de qualité et de maximiser le taux de réponse des répondants, chaque

série était composée d'uniquement 60 items au total. Un item étant composé d'une paire d'images, chacune étant associée à un court descriptif sous format textuel. Au total, ce sont donc 360 items (soit 720 choix de réponse unique) qui ont été testés afin de développer SWIPE. Ces items ont été construits par une équipe de 5 psychologues, sur base du corpus théorique et sémantique lié au modèle des Big Five.

Les répondants à ces séries ont été invités à l'étude car ils réunissaient 3 conditions essentielles : (1) avoir créé un compte AssessFirst en 2022, (2) avoir complété le questionnaire SHAPE en langue française, (3) avoir accepté les communications de prospection commerciale et scientifique d'AssessFirst dans le respect de la RGPD. Afin de maximiser le taux de réponse, plusieurs relances mail ont été effectuées.

Enfin, pour pouvoir étudier la qualité des items et la validité convergente de SWIPE au regard du BFI-2 et de son échelle d'humilité, les participants ayant complété une série de SWIPE ont également été invité à compléter la version française du BFI-2 (Lignier, Petot, Canada, Oliveira, Nicolas, Courtois, John, Plaisant & Soto, 2022), à laquelle nous avons également ajouté 12 items mesurant les 3 facettes du domaine d'humilité (Denissen, Soto, Geenen, John & Van Aken, 2022). Ce second questionnaire était donc composé de 72-items de type Likert, et accessible suite à la passation de SWIPE sur la plateforme Typeform.

En résumé, les participants ayant accepté de participer ont suivi la procédure suivante :

Étape 1
INVITATION

Réception du mail d'invitation présentant la démarche de collecte de données et ses finalités.

Étape 2
SWIPE

Complétion d'une série de SWIPE, composée de 60 items - environ 5 minutes de passation.

Étape 3
BFI-2 + H Scale

Complétion du BFI-2 + échelle d'humilité via la plateforme Typeform - environ 10 minutes de passation.

Au total, cette première phase a vu la participation de 2989 personnes - environ 500 répondants par série. Les descriptifs des échantillons de répondants sont présentés dans les tableaux 4.1 et 4.2 ci-après :

Série	Répondants	Genre			Expérience pro.
	Total	Femmes	Hommes	Non binaire	Moyenne
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. Descriptif du genre et de l'expérience moyenne des répondants pour la phase de séries.

Série	Niveau d'éducation					
	PhD	Master	Licence	Bac	Pro	Aucun
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. Descriptif du niveau d'éducation des répondants pour la phase de séries.

Ces quelques statistiques permettent ainsi de justifier de la diversité des utilisateurs qui ont participé à la première phase du développement de SWIPE. Si les résultats en termes de distribution des niveaux de diplôme ou d'éducation montrent une légère sur-représentativité des utilisateurs ayant un Master ou une Licence, celle-ci s'explique par deux raisons principales : (1) la représentativité naturelle et croissante des personnes ayant un niveau d'enseignement tertiaire dans la population - environ 50% selon les chiffres de l'OECD entre 2017 et 2020, et, (2) l'utilisation légèrement plus accrue de la solution AssessFirst pour le recrutement de profils cadres par nos clients - expliquant alors leur prévalence dans notre base de contacts.

4.2. Phase 2 : sélection des items

Les données récoltées à chaque série ont ensuite été analysées par nos équipes de psychométriciens, afin de choisir les items et choix de réponse qui mesuraient le mieux leur construit de référence. Aussi, ces items ont été sélectionnés sur base de plusieurs règles et conditions, notamment :

- Avoir une bonne corrélation avec le construit de référence ($r > .30 \mid r < -.30$) ;
- Que les deux choix de réponse d'un item remplissent la condition précédente ;
- Avoir une validité de contenu avec le construit optimale ;
- Sélectionner des items où la sémantique n'était pas trop répétitive pour une même facette ;
- Avoir un équilibre de choix de réponse « positifs » et « négatifs » pour une même facette ;
- Sélectionner des items diversifiés, l'idéal étant d'avoir un item pour chaque combinaison de 2 facettes ;
- Sélectionner des items où les personnages principaux sont diversifiés, en termes de genre et d'origine.

Ce processus de sélection nous a permis de sélectionner 135 items uniques sur les 360 testés au cours des 6 séries proposées. Ces 135 items ont ensuite été soumis à une nouvelle phase de collecte de données.

4.3. Phase 3 : série de validation

Afin de procéder à la sélection des items finaux qui composeront SWIPE, les 135 items choisis à l'étape précédente ont été consolidés au sein d'une série de « validation ». Celle-ci a pour objectif de collecter de l'information sur chaque item au sein d'un échantillon plus large. Les répondants à cette série finale ont été invités à l'étude car ils réunissaient 3 conditions essentielles : (1) avoir créé un compte AssessFirst en 2022, (2) avoir complété le questionnaire SHAPE en langue française, (3) avoir accepté les communications

de prospection commerciale et scientifique d'AssessFirst dans le respect de la RGPD. Afin de maximiser le taux de réponse, plusieurs relances mail ont été effectuées. De même, certains participants à la phase 1 ont été ré-invité à compléter la série finale. Ce recueil de données s'est déroulé du 09.01.23 au 10.02.23. Comme pour la première phase, les participants ont été invités à compléter SWIPE, puis l'inventaire BFI-2.

Série	Répondants	Genre			Expérience pro.
	Total	Femmes	Hommes	Non binaire	Moyenne
Validation	4457	54 %	33 %	13 %	10,9

Table 4.3. Descriptif du genre et de l'expérience moyenne des répondants de la série de validation.

Série	Niveau d'éducation					
	PhD	Master	Licence	Bac	Pro	Aucun
Validation	2 %	39 %	31 %	13 %	11 %	4 %

Table 4.4. Descriptif du niveau de diplôme des répondants de la série de validation.

Une nouvelle fois, les données récoltées ont ensuite été analysées par nos équipes de psychométriciens, afin de choisir les items qui mesuraient le mieux leur construit de référence. Un premier tri de ces items à été réalisé sur la base des mêmes critères que ceux définis lors de la phase 2. Ensuite, la sélection des items s'es faites sur d'autres critères qui ont permis de compléter les décision :

5. Version finale

75 La version finale de SWIPE est composée de 75 items à choix forcé. Parmi ces 75 items, 3 sont utilisés à des fins de recueil de données.

6 Nombre de traits de personnalité considérés par SWIPE. Ceux-ci permettent de couvrir l'ensemble de la personnalité.

18 Nombre de facettes de personnalité mesurées par SWIPE. Ces facettes sont directement issues du BFI-2 et de l'échelle ajoutée d'humilité.

5 Le temps de passation moyen. En moyenne, les répondants mettent 4,19 secondes pour répondre à un item.

6. Validité

Comment savoir si un questionnaire mesure réellement ce qu'il est censé mesuré ? Comment s'assurer de la bonne mesure de chaque échelle ainsi que la juste signification des résultats au questionnaire ? Les réponses à ces questions sont obtenues grâce à sa validation. L'objectif de la validation d'un questionnaire consiste à confirmer que celui-ci mesure réellement ce qu'il cherche à mesurer, et à quel point les résultats que nous pouvons en tirer son précis. Historiquement, la validité a été définie comme une corrélation entre un score à un questionnaire et un critère externe, qui mesure soit le même construit, ou qui mesure un construit censé être en lien avec le construit associé à ce score. Plusieurs types de validité sont nécessaires pour établir et assurer la validité d'un questionnaire. Les études de validité de SWIPE portent sur les types de validité suivants :

- **Validité de contenu** : comme son nom l'indique, la validité de contenu repose sur la nature sémantique du contenu de l'item par rapport au construit mesuré. En ce sens, le contenu doit être en rapport direct avec le construit qu'il est censé mesuré, et aussi couvrir tous les aspects principaux du construit mesuré ;
- **Validité de construit** : elle se réfère à la mesure dans laquelle le questionnaire évalue réellement la facette ou le construit psychologique qu'il est censé évaluer. Sont notamment proposées des analyses relatives à la saturation item-dimension, la corrélation inter-dimension, le RMSEA et les paramètres de distribution ;
- **Validité convergente** : la validité convergente fait référence au degré auquel deux mesures de construits qui devraient théoriquement être liés le sont. En d'autres termes, la validité convergente mesure à quel degré les résultats d'un questionnaire sont corrélés avec ceux d'un autre questionnaire qui évalue le même concept ;
- **Validité prédictive** : la validité prédictive d'un questionnaire de personnalité est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire de personnalité peuvent être utilisés pour prédire la performance future.

6.1. Validité de contenu

6.1.1. Introduction

La validité de contenu est liée à la pertinence du contenu du questionnaire. Elle vise à vérifier dans quelle mesure le questionnaire représente toutes les facettes d'un construit donné, et à quel point les items du questionnaire sont représentatifs du construit mesuré. Ce type de validité est important, au sens que la création d'items pour un questionnaire de personnalité reste avant tout un processus d'essais et erreurs (Tellegen & Waller, 2008). En ce sens, si elles sont mal développées, les échelles mesurées par un questionnaire peuvent contenir des items qui ne sont pas suffisamment représentatifs du contenu mesuré (Smith, Min, Ng, Haynes & Clark, 2022). La validité de contenu permet ainsi d'évaluer le degré auquel le contenu d'un item est lié au construit de personnalité qu'il est censé mesurer (Worthington & Whittaker, 2006; Colquitt, Sabey, Rodell & Hill, 2019). Historiquement, la validité de contenu ne repose pas sur une analyse statistique, mais sur une approche rationnelle pour lier le contenu de l'item au construit. Aussi, une démarche courante consiste à solliciter des juges experts, qui vont se prononcer quant à la pertinence des items, à travers un exercice manuel de classification des items. Plusieurs indicateurs sont alors calculés, comme l'accord inter-juge, qui représente la proportion de juges qui indiquent que l'item est sémantiquement bien lié au construit qu'il mesure (Anderson & Gerbing, 1991; Fleiss, 1981). Toutefois, cette approche est soumise à plusieurs limites, au sens qu'elle se révèle chronophage et cognitivement coûteuse (Krippendorff, 2018; Short, McKenny & Reid, 2018), et peut se voir influencée par la compétence des juges sélectionnés, qui peuvent se montrer imprécis dans leurs classifications (Fyffe, Lee & Kaplan, 2023).

Les techniques de Machine Learning (ML) et de NLP (Natural Language Processing), qui sont de plus en plus appliquées aux sciences comportementales, à la création et à l'analyse de contenus (Campion, Campion, Campion & Reider, 2016; Hommel, Wollang, Kotova, Zacher & Schmukle, 2022; Jiao & Lissitz, 2020; Lee, Fyffe, Son, Jia & Yao, 2023; Von Davier, 2018) permettent aujourd'hui de dépasser ces limites et de significativement optimiser le processus de validité de contenu. De récentes recherches (Fyffe, Lee & Kaplan, 2023) proposent ainsi une nouvelle approche, basée sur le NLP, mais utilisant des modèles de *transformers*. Les *transformers* sont un type de réseaux de neurones profonds qui sont utilisés pour convertir du texte en représentations numériques. Contrairement aux modèles de langage naturel précédents, qui utilisaient principalement des réseaux de neurones récurrents (RNN), les *transformers* se basent sur une architecture de traitement parallèle qui permet une meilleure prise en compte des relations entre les différents éléments de la séquence de texte. Aussi, en appliquant ces *transformers* à la classification de texte, les auteurs ont développé une approche automatisée de la validation de contenu des échelles de personnalité. Par rapport à l'approche traditionnelle précédemment décrite, cette méthode permet ainsi de réduire la lourdeur procédurale et cognitive, tout en optimisant les performances de classification. Aussi, cette méthodologie se pose comme une évolution majeure dans la capacité des éditeurs à construire des échelles de mesure efficaces, et à valider la qualité des contenus.

6.1.2. Comment ça marche ?

Les tâches de classification consistent à entraîner un modèle de classification à classifier un texte dans des catégories prédéfinies. Un modèle de classification est ainsi un type de modèle d'apprentissage automatique qui est utilisé pour prédire la classe ou la catégorie d'un objet ou d'une observation en fonction de ses caractéristiques : dans le cadre de la construction de SWIPE, il s'agit ainsi, en premier lieu d'entraîner un modèle qui sera en mesure de déterminer le trait de personnalité auquel se rattache chacun des items du questionnaire. Le développement de ce modèle de classification se déroule en 4 étapes principales :

- La création d'un jeu de données d'entraînement : pour construire un modèle de classification efficace, il est nécessaire de collecter un jeu de données qui contient des items de personnalité et le trait du Big Five auxquels ils appartiennent. Ces données doivent être représentatives des différentes classes que l'on souhaite prédire. Aussi, comme l'algorithme de classification apprend sur base de ces données, il est nécessaire de s'assurer qu'elles soient de bonne qualité, et puissent être utilisées ;
- La représentation textuelle : il s'agit de la façon dont les données textuelles (les items) sont encodées sous forme de vecteurs numériques qui peuvent être traités par des algorithmes d'apprentissage automatique. Les algorithmes d'apprentissage automatique ne peuvent pas traiter directement le texte brut, car ils travaillent avec des données numériques. Pour que les algorithmes d'apprentissage automatique puissent traiter les données textuelles, il est donc nécessaire de les encoder sous forme de vecteurs numériques. Cette représentation numérique permet de prendre en compte les caractéristiques importantes du texte, telles que les mots utilisés, leur ordre, leur fréquence, etc. ;
- L'entraînement du modèle : consiste à apprendre à un algorithme de classification à identifier les relations entre les caractéristiques d'entrée (les items) et la variable de sortie (le trait de personnalité à prédire). En d'autres termes, l'objectif de l'entraînement du modèle est de trouver une fonction qui lie les caractéristiques d'entrée à la classe de sortie ;
- L'évaluation du modèle : après avoir entraîné le modèle, il est évalué en utilisant un échantillon neutre. Différentes mesures de performance peuvent être utilisées pour évaluer la qualité du modèle, telle que l'exactitude (*accuracy*), la précision (*precision*), le rappel (*recall*) et le *F1-score*.

6.1.3. Le modèle de classification AssessFirst

Afin de capitaliser sur la richesse de la littérature scientifique et open-source quant à ce sujet, nos études se sont inscrites dans la continuité de certains résultats déjà obtenus par Fyffe, Lee et Kaplan (2023). Aussi, sur base de leurs résultats, nous avons fait plusieurs choix, permettant de répondre au mieux à notre objectif.

D'une part, si leurs études ont entraîné un modèle de classification à prédire les classes des 5 traits des Big Five, nos études pour SWIPE se doivent d'aller plus loin : en effet, au-delà des Big Five, SWIPE intègre également un domaine d'humilité. Nos équipes ont donc sélectionné un nouveau jeu de données d'entraînement, qui a été ajouté à celui déjà proposé et utilisé par Fyffe, Lee et Kaplan (2023), et qui comprenait notamment des items appartenant au trait d'humilité. Afin de nous assurer de la qualité des données et de ne pas pénaliser les performances du modèle, ces items ont été sélectionnés dans les bases open-sources (e.g., IPIP : International Personality Item Pool) et dans les questionnaires les plus réputés (e.g., BFI, BFI-2, BFI-10, HEXACO-100, HEXACO-60, HEXACO-24, BFAS, NEO-PI-R). L'enrichissement de cette base de données nous permettant ainsi d'également entraîner le modèle à prédire la classe d'humilité.

Trait	Nombre d'items
Extraversion	669
Agréabilité	762
Humilité	246
Ouverture	777
Concienciosité	780
Stabilité émotionnelle	671

Table 6.1. Nombre d'items d'entraînement par trait.

Note : nous faisons ici le choix d'une analyse de contenu par trait de personnalité, et non par facette. L'analyse par facette requiert d'être en mesure d'identifier et de recueillir suffisamment d'items d'entraînement structurés et de qualité par facette, afin de développer un modèle de classification efficace : en effet, les performances du modèle peuvent être drastiquement affectées en dessous de 40 exemples par classe (Fyffe, Lee & Kaplan, 2023). Au regard de la complexité à constituer cet échantillon d'entraînement, nous privilégions donc une analyse intermédiaire par trait, le temps de pouvoir créer un set d'entraînement par facette.

D'autre part, sur base des résultats obtenus par Fyffe, Lee et Kaplan (2023), nous avons fait le choix d'utiliser DeBERTa. DeBERTa - « Decoding-enhanced BERT with disentangled attention », est un modèle de traitement de langage naturel (NLP) basé sur un *transformer* (He, Liu, Tao & Chen, 2021). DeBERTa est une amélioration du modèle BERT (Bidirectional Encoder Representations from Transformers), qui est l'un des modèles de NLP les plus performants à ce jour. DeBERTa utilise une architecture *transformer* similaire à BERT, mais il présente plusieurs améliorations et innovations pour améliorer les performances sur différentes tâches de NLP, notamment : un décodage amélioré, une attention désentrelacée, une adaptation multi-tâches et une compression de modèle. À ce jour, DeBERTa a montré des performances exceptionnelles sur un large éventail de tâches de NLP, et ce y compris pour de la classification de textes.

6.1.4. Résultats et validité de contenu

Ce modèle de classification a ainsi appris à classifier efficacement des items de personnalité dans six traits, à partir d'un large corpus de questionnaires. En d'autres mots : sur base du contenu d'un item, ce modèle est maintenant en mesure de déterminer à quel trait de personnalité se rattache le plus cet item.

Tout l'intérêt est donc à présent d'utiliser ce modèle pour classifier les items de SWIPE dans les 6 traits de personnalité qu'ils sont censés mesurer, afin de déterminer ce qu'ils mesurent réellement et leur attribuer un trait. Les traits attribués par le modèle constituent ici le trait de « référence », au sens qu'il s'agit du trait duquel l'item est le plus représentatif. Ces labels attribués par notre modèle seront ensuite comparés aux traits que nous avons initialement attribués à chaque item de SWIPE, qui constituent quant à eux la « prédiction ». Aussi : (1) si nous avons classé les items dans le même trait que notre modèle, cela signifie qu'ils sont bien représentatifs du trait qu'ils mesurent, (2) si nous avons classés les items dans un autre trait que notre modèle, cela signifie qu'ils ne sont pas suffisamment représentatifs du trait qu'ils sont censés mesurer, ou qu'ils sont plus représentatifs d'un autre trait. À noter que même si nous prenons ici les résultats de classification du modèle comme source de référence en raison de ses performances, et au sens que les modèles automatisés de classifications d'items de personnalité s'avèrent souvent plus performants que les juges humains (83% d'exactitude pour DeBERTa, 71% pour un juge humain, voir Fyffe, Lee et Kaplan, 2023), il serait toutefois naïf de ne pas garder en tête cette marge d'erreur, qui appelle à garder un regard critique sur les résultats.

Pour évaluer la validité de contenu des items de chaque trait, 4 indicateurs sont ici mesurés :

- L'*accuracy* : mesure la proportion de prédictions correctes par rapport à toutes les prédictions effectuées. Il s'agit donc de la capacité à prédire correctement les observations positives et négatives. Elle se calcule en divisant le nombre total de prédictions correctes par le nombre total de prédictions effectuées. L'*accuracy* varie de 0 à 1 ;

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

- La *precision* : mesure la proportion de prédictions positives qui sont correctes parmi toutes les prédictions positives. En d'autres termes, la *precision* mesure notre capacité à ne pas classer à tort une observation négative comme positive. Elle se calcule en divisant le nombre de prédiction positives correctes par le nombre total de prédictions positives. La *precision* varie de 0 à 1 ;

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

- Le *recall*, ou *sensibilité* : mesure la proportion de vrais exemples positifs qui sont correctement prédits parmi tous les exemples positifs. En d'autres termes, le *recall* mesure la capacité à trouver toutes les observations positives. Il se calcule en divisant le nombre de prédictions positives correctes par le nombre total d'exemples positifs. Le *recall* varie de 0 à 1 ;

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

- Le *F1-score* : mesure combinée de la précision et du recall. Il s'agit de la moyenne harmonique de la précision et du recall. Le F1-score peut être considéré comme l'indicateur global d'efficacité. Le F1-score varie de 0 à 1, où une valeur de 1 indique une performance optimale en termes de *precision* et de *recall* ;

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

D'un point de vue général, il n'y a pas de scores universellement bons ou mauvais pour chacune de ces mesures. Les scores dépendent avant tout du contexte et des exigences spécifiques du problème de classification. Par exemple, dans certaines applications telles que la détection de fraudes financières, il peut être crucial d'avoir une haute *precision* pour minimiser le nombre de faux positifs, même si cela peut réduire le *recall* et manquer certains cas de fraude. Dans d'autres applications, telles que la détection de spam dans les e-mails, un *recall* élevé peut être plus important pour s'assurer que tous les messages de spam sont identifiés, même si cela peut augmenter le nombre de faux positifs. Néanmoins, il est généralement admis que, pour chacun de ces indicateurs, et notamment pour le *F1-score* : (1) un score ⩾ 0.9 est excellent, (2) un score compris en 0.8 et 0.9 est bon, (3) un score compris entre 0.7 et 0.8 est satisfaisant, (4) un score entre 0.5 et 0.7 est passable, (5) un score ⩽ 0.5 est considéré comme très insuffisant.

Les résultats obtenus suite au test des items SWIPE sont présentés dans le tableau 6.2.

Traits	Accuracy	Precision	Recall	F1-score
Extraversion	.75	.87	.75	**.81**
Agréabilité	.89	1	.89	**.94**
Humilité	1	.70	1	**.82**
Ouverture	.96	1	.96	**.98**
Concienciosité	.93	.96	.93	**.95**
Stabilité émotionnelle	.96	.92	.96	**.94**
	.92	*.91*	*.92*	*.91*

Table 6.2. Performance en fonction du trait de personnalité (SWIPE).

6.1.5. Interprétation des résultats

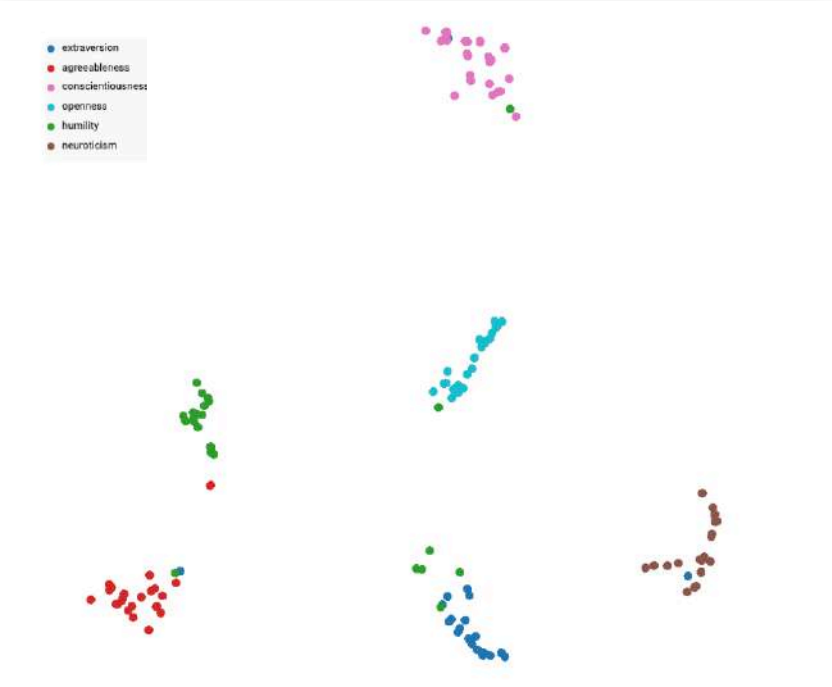
Si ce type d'analyse est nouveau, et constitue très certainement une première dans l'étude de la validité de contenu pour le développement d'un nouveau questionnaire de personnalité, les résultats présentés apportent des informations intéressantes quant à la qualité des items de SWIPE, et leur représentativité des traits de personnalité qu'ils sont supposés mesurer. En ce sens :

- Pour le trait « Agréabilité », les résultats sont excellents, avec une *précision* de 1, un *recall* de .89 et un *F1-score* de .94. La majeure partie des items appartenant au trait Agréabilité (comme défini par notre modèle) ont correctement été classifiés lors du développement de SWIPE, et tous les items classifiés dans ce trait par SWIPE le sont également par notre modèle de référence ;
- Pour le trait « Ouverture », les résultats sont excellents, avec une *précision* de 1, un *recall* de .96 et un *F1-score* de .98. La majeure partie des items appartenant au trait Ouverture (comme défini par notre modèle) ont correctement été classifiés lors du développement de SWIPE, et tous les items classifiés dans ce trait par SWIPE le sont également par notre modèle de référence ;
- Pour le trait « Conscienciosité », les résultats sont excellents, avec une *précision* de .96, un *recall* de .93 et un *F1-score* de .95. La majeure partie des items appartenant au trait Conscienciosité (comme défini par notre modèle) ont correctement été classifiés lors du développement de SWIPE, et tous les items classifiés dans ce trait par SWIPE le sont également par notre modèle de référence ;
- Pour le trait « Stabilité émotionnelle », les résultats sont excellents, avec une *précision* de .92, un *recall* de .96 et un *F1-score* de .94. La majeure partie des items appartenant au trait Stabilité émotionnelle (comme défini par notre modèle) ont correctement été classifiés lors du développement de SWIPE, et tous les items classifiés dans ce trait par SWIPE le sont également par notre modèle ;

- Pour le trait « Extraversion », on remarque une *precision* élevée (.87) mais un *recall* plus faible (.75), ce qui signifie que, globalement, les items classifiés en Extraversion par SWIPE le sont également par notre modèle de référence, mais que certains items qui sont davantage représentatifs du trait Extraversion (comme défini par notre modèle) ont été classifiés dans un autre trait lors du développement de SWIPE ;

- Pour le domaine « Humilité », la *précision* est plus faible (.70) et le *recall* est beaucoup plus élevé (1), ce qui signifie que tous les items représentatifs du trait Humilité, tel que défini par notre modèle, ont bien été attribué à ce trait lors du développement de SWIPE, mais que nous avons trait attribué certains items au trait Humilité, alors qu'ils semblent plus représentatif d'un autre domaine.

Si les résultats présentés sont excellents pour 4 traits, il convient de davantage s'intéresser aux traits Extraversion (F1-score = .81) et Humilité (F1-score = .82) qui, s'ils présentent de bons résultats, sont très légèrement en retrait. Aussi, une analyse approfondie des items ayant fait l'objet d'une classification divergente permet de trouver une explication conceptuelle. En effet, la plupart des items initialement classés en Humilité lors du développement de SWIPE, sont soient identifiés par notre modèle comme plus représentatif du trait Extraversion, ou du trait Agréabilité. Ce pattern de classification prend probablement source des liens naturels - et démontrés, qui existent entre ces domaines de personnalité. En ce sens, les facettes « assertiveness » et « sociability », appartenant au trait Extraversion, présentent des corrélations négatives avec le trait Humilité (Lee, Ashton & De Vries, 2022; Ludeke, Bainbridge, Liu, Zhao, Smillie & Zettler, 2019), tandis que le trait Agréabilité est quant à lui positivement corrélé (Lee, Ashton & De Vries, 2022). De même, le trait Humilité est très fortement en lien avec la Dark Triad (Howard & Van Zandt, 2020), elle même extrêmement corrélée à l'assertiveness (Kaufman, Yaden, Hyde & Tsukayama, 2019). Enfin, nos propres études sur un échantillon de participants ayant complété le BFI-2 montrent des corrélations significatives entre plusieurs facettes liées à ces traits de personnalité, notamment entre assertiveness ~ modesty ($r = -.36$; $p < 2.2e\text{-}16$), entre respectfulness ~ modesty ($r = .39$; $p < 2.2e\text{-}16$) ou encore entre respectfulness ~ sincerity ($r = -.35$; $p < 2.2e\text{-}16$). Ces constats sont d'ailleurs confirmés par les corrélations inter-dimensions obtenues dans la littérature (Soto & John, 2017). Au regard des liens existants entre ces facettes et traits de personnalité, il n'est donc pas incohérent de voir des items qui sont classifiés différemment entre le modèle de langage et le modèle SWIPE.



Graphique 6.1. Clusters de classification des items SWIPE par le modèle NLP.

Cette représentation spatiale permet de mieux visualiser les quelques erreurs de classification précédemment citées. En ce sens, certains items initialement classés dans le trait Humilité (points verts sur le graphique) ont été classés dans dans les traits Extraversion (points bleu foncé sur le graphique) ou Agréabilité (points rouges sur le graphique). Cette analyse par cluster confirme ainsi les liens étroits entre ces trois traits de personnalité (Lee, Ashton & De Vries, 2022; Ludeke, Bainbridge, Liu, Zhao, Smillie & Zettler, 2019; Lee, Ashton & De Vries, 2022; Soto & John, 2017). Ces liens sont aussi visualisés par la proximité spatiale des 3 traits sur le graphique.

À l'inverse, les 3 autres traits, liés à l'Ouverture (bleu clair), la Conscienciosité (rose) et la Stabilité émotionnelle (marron) sont représentés par des clusters qui sont visuellement plus indépendants.

Aussi, au-delà d'interroger les résultats et la qualité des items de SWIPE, cette conclusion ouvre plutôt au besoin de considérer les biais potentiellement inhérents au modèle de classification entraîné. En effet, ce modèle a été entraîné en partant du principe qu'un item ne pouvait mesurer qu'un seul et unique trait de personnalité (on parle de classification multiclasse ou *multiclass classification*). Toutefois, de plus en plus, les éditeurs de questionnaires de personnalité se tournent vers l'usage de « blended items » qui mesurent différentes facettes de personnalité (Schwaba, Rhemtulla, Hopwood & Bleidorn, 2020), comme c'est le cas dans le questionnaire SWIPE. Il est donc nécessaire de s'intéresser à des techniques de classification multilabel (ou *multilabel classification*) qui consistent à attribuer plusieurs labels à une seule observation (Fyffe, Lee et Kaplan, 2023) - cela signifiant que chaque observation, ou chaque item dans notre cas précis, peut appartenir à plus d'une catégorie simultanément. En général, la classification multilabel est plus complexe que la classification multiclasse, car elle nécessite une analyse plus fine de chaque observation pour déterminer les différentes étiquettes qui lui correspondent. Les algorithmes de classification multilabel peuvent être plus coûteux en termes de temps de calcul et de puissance de traitement, mais ils sont souvent plus flexibles et peuvent être plus adaptés à certaines tâches. Aussi, nos études quant à ce sujet sont en cours, et seront ajoutées à ce guide lorsqu'elles seront disponibles.

6.1.6. Comparaison à d'autres questionnaires

Pour aller plus loin dans la compréhension des résultats et de la qualité des items SWIPE, nous pouvons également comparer les indicateurs obtenus pour SWIPE à ceux obtenus pour d'autres questionnaires de personnalité. Ces analyses seront ajoutées à ce manuel quand elles seront disponibles.

6.1.7. Conclusion

Les résultats démontrent une excellente validité de contenu des échelles de SWIPE au niveau des traits. En effet, les indicateurs mesurés mettent en évidence que les items de SWIPE sont bien représentatifs des traits de personnalité qu'ils mesurent. Les résultats sont excellents pour 4 traits parmi les 6 mesurés (Agréabilité, Ouverture, Conscienciosité et Stabilité émotionnelle), et bons pour les deux autres (Extraversion et Humilité). Aussi, bien que les résultats concernant les deux derniers traits cités restent bons, il convient de mentionner qu'ils sont, qui plus est, négativement impacté par un chevauchement conceptuel entre les deux traits, et par le fait de l'entraînement par classification multiclasse de notre modèle. En somme, les résultats de validité de contenu ici présentés attestent du bien-fondé théorique, conceptuel et sémantique de SWIPE, et que ses contenus sont représentatifs des Big Five et de l'échelle ajoutée d'humilité.

6.2. Validité de construit

6.2.1. Introduction

La validité de construit permet de savoir si le test mesure réellement le construit théorique souhaité, ou autre chose. Ce type de validité chevauche certains des autres aspects de la validité, car toute preuve de validité contribue à la compréhension de la validité de construit d'un instrument d'évaluation. La validité de construit est importante, car elle influe sur l'interprétation des scores d'un questionnaire. Si un questionnaire prétend mesurer une facette de personnalité spécifique, comment être sûr qu'il mesure réellement cette facette ? Si un questionnaire est censé mesurer une facette spécifique, mais qu'en réalité, ce n'est pas le cas, toute interprétation du score est incorrecte et pourrait conduire à des décisions biaisées. La validité de construit ne cherche pas simplement à savoir si un questionnaire mesure une facette. Aussi, elle considère des investigations complexes cherchant à savoir si les interprétations des résultats des tests sont cohérentes avec un réseau impliquant des termes théoriques et d'observation (Cronbach & Meehl, 1955).

Il n'existe pas de méthode unique pour déterminer la validité de construit, mais différentes méthodes et approches sont combinées. Pour évaluer la validité de construit de SWIPE, nous privilégions ici quatre méthodes complémentaires, à savoir la saturation item-dimension et la corrélation inter-dimensions (Thurstone, 1947; Bollen, 1989; McDonald, 2013), le RMSEA et enfin la présentation des paramètres de distribution (Fisher, 1912, 1920, 1921, 1922) :

- la **saturation item-dimension** fait référence à la corrélation entre les scores d'un item et les scores totaux de la dimension ou du facteur auquel il est censé appartenir. En d'autres termes, si un item est censé mesurer une dimension spécifique, alors il devrait être étroitement associé aux autres items qui mesurent cette dimension. Ainsi, plus la corrélation entre un item et la dimension est élevée, plus l'item est considéré comme étant fortement lié à cette dimension et donc plus valide. Pour la saturation item-dimension, une valeur supérieure ou égale .40 est généralement considérée comme satisfaisante et adéquate : cela suggère que l'item mesure bien la dimension à laquelle il est censé se rapporter (Campbell & Fiske, 1959; Nunnally, 1978; Hair, Black, Babin & Anderson, 1998). Une saturation inférieure à .40 peut être acceptable si elle est soutenue par une justification théorique ;

- la **corrélation inter-dimensions** évalue la relation entre les scores de différents facteurs ou dimensions mesurées par un test. Si deux dimensions sont censées être distinctes et indépendantes, alors elles devraient avoir des scores faiblement corrélés. En revanche, si les dimensions sont étroitement liées ou si elles se chevauchent, les scores devraient être plus fortement corrélés. Il n'existe pas de seuil universel pour la corrélation inter-dimension. Toutefois, certains auteurs suggèrent que la corrélation inter-dimensions ne devrait pas dépasser la racine carrée de la variance de chaque dimension pour établir la validité discriminante (Hair, Black, Babin & Anderson, 1998). Il est généralement souhaitable que les dimensions soient relativement indépendantes, bien qu'il puisse exister certaines corrélations modérées entre les dimensions qui sont justifiées par le modèle théorique sous-jacent. Si les corrélations entre les dimensions ne correspondent pas aux attentes théoriques, cela peut indiquer un problème de validité de construit. Il est donc nécessaire de pouvoir comparer ces corrélations à celles du construit théorique fondateur et de référence (le Big Five Inventory-2 dans le cas du questionnaire SWIPE) ;

- **le RMSEA**, ou Root Mean Square Error of Approximation, est une mesure d'ajustement qui est souvent utilisée pour évaluer l'adéquation d'un modèle de mesure. Le RMSEA mesure la différence entre les données observées et les données ajustées du modèle, corrigée pour le nombre de paramètres libres du modèle. Plus précisément, le RMSEA évalue l'ajustement absolu du modèle en comparant la variance non expliquée dans les données avec la variance non expliquée attendue dans les données selon le modèle. Généralement, un RMSEA < .05 indique un bon ajustement du modèle aux données (Steiger & Lind, 1980; Browne & Cudeck, 1993) ;

- **les paramètres de distribution** correspondent à la distribution des scores dans les questionnaires. Ils permettent d'identifier les scores atypiques, d'explorer les différences individuelles dans la distribution des scores, et de mieux interpréter les résultats.

6.2.2. Saturation item-dimension

Les dernières études de saturation item-dimension de SWIPE ont été conduites en avril 2023 (N = 4457) sur les items principaux mesurant chaque dimension. Le tableau ci-dessous présente les résultats de cette analyse : (1) la saturation varie de .30 ≤ r ≤ .68, (2) la saturation moyenne par facette varie de .46 ≤ r ≤ .58. Ces conclusions permettent ainsi d'attester de saturations item-dimension satisfaisantes et adéquates.

Traits	Facettes	i1	i2	i3	i4	i5	i6	i7	i8	Moyenne
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**
AGRÉABILITÉ	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**
HUMILITÉ	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**
OUVERTURE	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**
CONSCIENCIOSITÉ	Organization	.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**
STABILITÉ ÉMOTIONNELLE	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. Saturation item-dimension de SWIPE.

6.2.3. Corrélation inter-dimension

Les dernières études de corrélation inter-dimension de SWIPE ont été conduites en avril 2023 (N = 4457). Afin de pouvoir étudier la dynamique de ces corrélations inter-dimension au regard du modèle théorique sous-jacent et valider leur cohérence, les analyses suivantes sont proposées : (1) étude des corrélations inter-dimension de SWIPE - voir tableau 6.5, (2) étude des corrélations inter-dimension du BFI-2 - voir tableau 6.6, (3) analyse de cohérence entre les deux matrices de corrélations avec la corrélation de rang de Spearman ou ρ de Spearman, (4) analyse de l'effet de taille avec le q de Cohen - voir tableau 6.7, (5) une revue théorique explicative des liens entre certaines facettes est également proposée - voir tableau 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. Corrélations inter-dimension de SWIPE.

Dans l'ensemble, les dimensions sont faiblement corrélées, ce qui permet de justifier d'un niveau de consistance acceptable. La comparaison aux corrélations inter-dimension du BFI-2 nous permet aussi d'étayer l'analyse en analysant ces corrélations au regard du modèle théorique sous-jacent.

	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. Corrélations inter-dimension du BFI-2.

Afin de nous assurer de la cohérence entre ces deux matrices, nous utilisons ici le ρ de Spearman. Il s'agit d'une mesure de corrélation non-paramétrique entre deux variables. Contrairement à la corrélation de Pearson, qui mesure la relation linéaire entre deux variables continues, la corrélation de Spearman évalue la relation monotone entre deux variables. La corrélation de Spearman utilise les rangs des observations de chaque variable plutôt que leurs valeurs réelles pour calculer la corrélation. Les observations sont classées par ordre croissant ou décroissant selon leur valeur, et les rangs correspondants sont assignés. La corrélation de Spearman varie de -1 à 1, où une valeur de 1 indique une relation monotone positive parfaite, une valeur de -1 indique une relation monotone négative parfaite, et une valeur de 0 indique l'absence de relation monotone entre les variables.

Les résultats montrent une valeur ρ = .66 ($p < 2.2e\text{-}16$), permettant de démontrer la cohérence entre les deux matrices de corrélation. En somme, les corrélations inter-dimension présentent dans SWIPE se retrouve également dans le construit théorique fondateur, et sont ainsi inhérentes aux facettes et concepts mesurés.

Pour aller plus loin dans la compréhension des convergences et divergences entre les deux matrices, nous proposons aussi une analyse des effets de taille avec le q de Cohen. Le coefficient de Cohen, est une mesure de l'effet de la taille d'une différence entre deux groupes dans une étude statistique. Le coefficient de Cohen varie de -1 à 1, où 0 indique qu'il n'y a pas de différence entre les groupes, 1 indique une différence maximale et -1 indique une différence maximale dans l'autre sens. En général, une valeur $q \approx .0$ indique qu'il n'y a aucune différence, une valeur $q \approx .3$ correspond à une différence faible, $q \approx .5$ correspond à une différence moyenne, et $q \approx .8$ correspond à une différence forte. Aussi, pour nous assurer de la cohérence des corrélations inter-dimension entre les deux matrices, nous cherchons à obtenir des valeurs de q au plus proche de 0. Les résultats de cette analyse sont présentées dans le tableau ci-dessous.

	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. Analyse des coefficients de Cohen.

En conclusion, les corrélations inter-dimension de SWIPE sont relativement faibles et correspondent aux attentes théoriques : en ce sens, il n'existe pas de différence significative entre les corrélations inter-dimension mises en lumière dans SWIPE, et celles du BFI-2. Les deux matrices sont équivalentes. Aussi, les corrélations inter-dimensions les plus fortes concernent des facettes dont les liens ont à de multiples reprises étaient identifiés et justifiés dans la littérature scientifique :

Facette n°1	Facette n°2	Études qui soutiennent l'existence d'un lien
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	Organization	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. Revue théorique des liens inter-dimensions.

6.2.4. RMSEA

Les dernières études de RMSEA pour SWIPE ont été conduites en avril 2023. Pour chaque facette, l'indice RMSEA est inférieur à .05 (mean RMSEA = .01), indiquant une bonne adéquation du modèle aux données. Cela suggère que le modèle a une bonne capacité à expliquer les relations entre les variables mesurées et que les différences entre les données observées et les données prédites par le modèle sont faibles. Ces informations apportent une preuve de validité de construit supplémentaire au questionnaire SWIPE.

Facettes	RMSEA	Facettes	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. Indice RMSEA pour chaque facette.

6.2.5. Paramètres de distribution

La distribution des scores obtenus par un questionnaire de personnalité est un aspect important de la validité de construit du questionnaire. En effet, la manière dont les scores sont distribués pour chaque facette de personnalité peut fournir des informations importantes sur la manière dont le test mesure réellement cette facette, ainsi que sur la manière dont les scores sont interprétés. Nos analyses de paramètres de distribution se centrent sur 6 paramètres principaux et nécessaires :

- la **moyenne**, qui est un indicateur de la tendance centrale des scores dans la distribution ;
- la **médiane**, qui est la valeur qui divise la distribution en deux parties égales : la moitié des scores sont supérieurs à la médiane, et l'autre moitié sont inférieurs. C'est également un indicateur de la tendance centrale des scores dans la distribution, qui est moins sensible aux scores extrêmes que la moyenne ;
- l'**écart-type**, qui est une mesure de la dispersion des scores autour de la moyenne. Il est calculé en prenant la racine carrée de la variance des scores ;
- **l'étendue** des scores ;
- l'**asymétrie** , qui est calculée en comparant la fréquence des scores à gauche et à droite de la moyenne. Si la distribution est parfaitement symétrique, l'asymétrie est nulle. Si la distribution est asymétrique vers la gauche, l'asymétrie est négative. Si la distribution est asymétrique vers la droite, l'asymétrie est positive ;
- l'**aplatissement** (ou coefficient d'aplatissement), qui est calculé en comparant la distribution des scores à une distribution normale. Si la distribution est moins aplatie que la distribution normale, l'aplatissement est négatif. Si la distribution est plus aplatie que la distribution normale, l'aplatissement est positif.

Les attentes pour les paramètres de distribution dépendent du contexte et de l'instrument de mesure utilisé. Toutefois, en général, voici ce que l'on attend pour de « bons » paramètres de distribution :

- La moyenne doit être proche de la valeur médiane : cela indique que la distribution est symétrique. Si la moyenne est significativement différente de la médiane, cela peut indiquer une asymétrie de distribution ;
- L'écart-type doit être raisonnable et suffisamment grand pour capturer les différences individuelles dans la dimension mesurée, mais pas trop grand pour ne pas diluer les différences entre les individus. En général, on s'attend à ce que l'écart-type soit d'environ 2 pour les échelles de personnalité en 10 points ;
- L'étendue doit capturer la variation dans la dimension mesurée, mais ne pas être trop grande pour diluer les différences entre les individus. En général, on s'attend à ce que l'étude soit comprise entre 4 et 6 ;
- L'asymétrie (skewness) doit être proche de 0 (distribution symétrique). Si l'asymétrie est significativement différente de 0, cela peut indiquer une asymétrie dans la distribution ;
- L'aplatissement (kurtosis) doit être proche de 0 (distribution normale). Si l'aplatissement est significativement différent de 0, c'est que la distribution est soit plus plate, soit plus pointue.

Il est important de noter que ces attentes peuvent varier en fonction du contexte et de l'instrument de mesure utilisé. Par exemple, pour certains questionnaires de personnalité, il peut être normal d'avoir une distribution asymétrique ou une étendue plus grande. En ce sens, il est nécessaire de mettre en perspective les résultats ci-après présentés pour SWIPE avec ceux généralement obtenus dans littérature scientifique avec le BFI. Aussi, plusieurs recherches ont démontré des aplatissements légèrement plus négatifs pour le BFI (Plaisant, Courtois, Réveillère, Mendelsohn & John, 2009; Rammstedt, 2007; DeYoung, Carey, Krueger & Ross, 2016). Par exemple, une étude menée par Plaisant, Courtois, Réveillère, Mendelsohn et John (2009), relative à la validation factorielle des BFI en langue française, a montré un aplatissement légèrement négatif - entre -.53 et .56. Il convient aussi de noter que de légères asymétries et aplatissements ne sont pas inhabituels dans les mesures des facettes de personnalité, et ne remettent pas en question la validité et la fidélité du BFI-2. Les distributions des scores de personnalité sont effet souvent légèrement asymétriques ou avec des coefficients d'aplatissement différents de zéro.

Les dernières études de paramètres de distribution de SWIPE ont été conduites en avril 2023 (N = 4457).

Facettes	Moyenne	Médiane	Écart-type	Étendue	Asymétrie	Aplatissement
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. Paramètres de distribution de SWIPE.

Les paramètres de distribution sont ainsi cohérents avec les standards attendus, et également avec la littérature scientifique relative au modèle théorique sous-jacent, le BFI-2. La normalité des distributions peut être interprétée par des indices comme la superposition des moyennes et des médianes, et à l'aide des coefficients de symétrie et d'aplatissement qui sont proches de 0.

6.2.6. Conclusion

Plusieurs résultats ont permis de démontrer la validité de construit de SWIPE, et notamment : (1) les saturations inter-dimensions répondent aux standards requis, (2) les corrélations inter-dimensions sont globalement faibles, et les corrélations les plus fortes convergent avec celles inhérentes aux construits mesurés, et trouvent un support marqué dans la littérature scientifique, (3) les indices RMSEA pour chaque facette respectent largement les seuils adéquats, et enfin, (4) les paramètres de distribution sont bons et reflètent ceux théoriquement attendus. SWIPE mesure donc bien les facettes qu'il prétend évaluer.

6.3. Validité convergente

6.3.1. Introduction

La validité convergente est une mesure de la similitude entre les scores d'un test de personnalité et ceux d'autres tests ou mesures qui évaluent la même dimension ou facteur de personnalité. En d'autres termes, elle permet de vérifier si un test de personnalité mesure effectivement ce qu'il est censé mesurer. Plus spécifiquement, la validité convergente s'intéresse à la corrélation entre les scores d'un test et ceux d'autres mesures ou tests qui évaluent la même facette de personnalité. Une forte corrélation entre les scores indique que les échelles sont convergentes et mesurent le même construit, ce qui renforce la validité du test.

Il convient de noter qu'il n'existe pas de seuil « officiel » pour juger de la qualité de la convergence entre les deux mesures. Aussi, le seuil approprié dépend du contexte spécifique dans lequel le questionnaire est utilisé et des caractéristiques de la population cible. De plus, la validité convergente doit être évaluée en conjonction avec d'autres mesures de validité pour avoir une évaluation complète de la qualité du test de personnalité. Cependant, plusieurs auteurs et recherches ont proposé quelques suggestions ou seuils de satisfaction : (1) une corrélation de .7 ou plus entre un questionnaire et d'autres mesures qui évaluent la même dimension est un indicateur d'une très forte validité convergente par Campbell et Fiske (1959), (2) une corrélation de .6 est recommandée comme seuil de validité par Worthington et Whittaker (2006), (3) une corrélation de .5 ou plus est considérée comment une bonne validité convergente par Bagozzi et Yi (1988) et par Revelle et Condon (2015), (4) une corrélation de .4 est considérée comme acceptable par Nunnally et Bernstein (1994). En somme, s'il n'existe pas de consensus clair ou de règle d'or (Marsh, Hau & Wen, 2004) sur la valeur exacte à utiliser comme seuil de validité convergente, il est recommandé de viser des corrélations de .5 minimum pour justifier d'une bonne validité convergente d'un questionnaire de personnalité.

6.3.2. Validité convergente avec le BFI-2

Nous étudions la validité convergente du questionnaire SWIPE, avec la version française du BFI-2 (Lignier, Petot, Canada, Oliveira, Nicolas, Courtois, John, Plaisant & Soto, 2022). Les dernières études de validité convergente de SWIPE avec le BFI-2 ont été conduites en avril 2023 (N = 4457).

Facette SWIPE	Facette BFI-2	*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. Validité convergente (*r*) des facettes mesurées par SWIPE avec le BFI-2.

6.3.3. Conclusion

Les analyses présentées démontrent une bonne validité convergente de SWIPE avec le BFI-2. Les corrélations sont comprises entre $.55 \leq r \leq .77$. En ce sens, les corrélations de 9 facettes dépassent le seuil de .7 proposé par Campbell et Fiske (1959), et toutes les corrélations dépassent les seuils de .5 proposés par d'autres. Nous pouvons dès lors conclure que les relations entre SWIPE et le BFI-2 sont suffisamment fortes pour valider une base de construits similaires.

6.4. Validité prédictive

La validité prédictive d'un questionnaire de personnalité est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire de personnalité peuvent être utilisés pour prédire la performance future. La preuve de validité prédictive est particulièrement applicable lorsque l'on souhaite faire une inférence, à partir d'un score du

questionnaire, de la position de l'évalué sur une autre variable de critère évaluée de manière indépendante à une date ultérieure. Les études de validité prédictive liées à SWIPE sont actuellement en cours de réalisation.

6.5. Conclusion

La validation d'un questionnaire de personnalité est cruciale pour s'assurer que les mesures obtenues sont précises. Dans cette étude, nous avons examiné la validité de contenu, la validité de construit, et la validité convergente de SWIPE. Nos analyses montrent que : (1) SWIPE couvre les construits théoriques qu'il est censé mesurer, (2) le questionnaire est bien structuré et montre une bonne homogénéité de la mesure, (3) il existe une forte corrélation entre SWIPE et le BFI-2, ce qui confirme la similitude des construits mesurés. Dans l'ensemble, les résultats obtenus répondent aux standards psychométriques les plus exigeants, et démontrent la validité de SWIPE : SWIPE mesure bien les facettes de personnalité présentées. Aller plus loin dans les qualités psychométriques de SWIPE requiert toutefois de s'intéresser à sa fidélité. En ce sens, un questionnaire doit être valide et fidèle pour être utilisé dans le cadre de décisions professionnelles (recrutement, mobilité, etc.). En effet, un questionnaire valide mais non fidèle signifierait que le test mesure bien ce qu'il doit mesurer, mais que les scores individuels sont incohérents. Au contraire, un questionnaire valide et fidèle signifie que le questionnaire mesure systématiquement ce qu'il est censé mesurer : en d'autres mots, il frappe constamment dans le mille. Les preuves de fidélité sont présentées dans le chapitre suivant.

Synthèse de validité

L'objectif de la validation d'un questionnaire consiste à confirmer que celui-ci mesure réellement ce qu'il cherche à mesurer, et à quel point les résultats que nous pouvons en tirer son précis. Les études portent sur les validités de contenu, de construit, convergente et prédictive.

.91 F1-Score moyen. Les résultats démontrent une excellente validité de contenu des échelles de SWIPE au niveau des traits.

.01 RMSEA moyen, indiquant une bonne adéquation du modèle aux données, et apportant des preuves de validité de construit du questionnaire SWIPE.

.70 Corrélation moyenne avec les échelles du BFI-2, qui démontre la validité convergente de SWIPE, et valide la mesure de construits similaires.

7. Fidélité

Comment savoir si les résultats d'un questionnaire sont fiables ? Comment s'assurer que le questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ? Les réponses à ces questions sont obtenues grâce à l'étude de sa fidélité. Alors que la validité renseigne sur la capacité d'un questionnaire à réellement mesurer ce que l'on souhaite mesurer, la fidélité permet de savoir si cette mesure sera consistante et fiable à chaque fois que ce même questionnaire est complété par une même personne. En somme, la fidélité d'un questionnaire est une mesure de sa cohérence ou de sa stabilité dans le temps. Elle vise à déterminer si un questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ou à des personnes similaires. L'objectif de la fidélité est donc de garantir que les résultats obtenus sont fiables et précis. La fidélité d'un questionnaire peut s'évaluer de deux manières différentes et complémentaires :

- **La cohérence interne,** qui est une mesure statistique utilisée pour évaluer la fidélité d'un test psychométrique. Elle évalue l'homogénéité ou la similarité des différents items d'un test qui sont supposés mesurer la même dimension psychologique. Autrement dit, la cohérence interne évalue si plusieurs éléments qui proposent de mesurer la même chose produisent des scores similaires ;

- **La fidélité test-retest,** qui est une méthode pour évaluer la fidélité d'une mesure en mesurant la même variable à deux moments différents. Elle permet de mesurer la stabilité temporelle de la mesure et d'estimer la proportion de la variance totale qui est attribuable à l'erreur de mesure. Elle est souvent utilisée dans les études longitudinales ou pour évaluer la stabilité d'un test sur une période de temps donnée.

7.1. Cohérence interne

Le concept de cohérence interne a été introduit par le psychologue Lee Cronbach dans les années 1950. Ce dernier propose l'alpha de Cronbach comme une mesure de la fidélité d'un questionnaire, qui calcule la corrélation moyenne entre les différents items. L'alpha de Cronbach a depuis été largement utilisé comme mesure de la cohérence interne dans les tests psychométriques. Si l'alpha de Cronbach a gagné en popularité en raison de sa facilité de calcul et d'interprétation, il présente toutefois plusieurs limites dans l'évaluation de la fidélité de questionnaires plus modernes : (1) il est difficile d'obtenir des consistances internes élevées dans les questionnaires à choix forcé, car ce format fausse la cohérence interne des instruments (Brown & Maydeu-Olivares, 2013), (2) l'alpha de Cronbach a tendance à sous-estimer la fidélité (Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019), (3) il est davantage adapté aux échelles unidimensionnelles - où chaque item ne mesure qu'une seule facette (Cortina, 1993), et, (4) il est fortement influencé par le nombre d'items, le nombre de dimensions orthogonales et la moyenne des corrélations entre les items (Cortina, 1993). Son usage est donc de plus en plus critiqué, et non recommandé.

Pour dépasser ces limites, plusieurs auteurs recommandent l'utilisation d'un autre indicateur, l'Omega de McDonald, introduit par J.B. McDonald en 1970. McDonald, un psychologue américain, a créé cette mesure de fidélité comme alternative à l'alpha de Cronbach, en partant d'une approche factorielle. L'omega présente notamment deux avantages : (1) il tient compte de la force de l'association entre les items et un construit, (2) il tient compte du lien entre les items et l'erreur de mesure. Depuis sa création, l'omega de McDonald a été largement utilisé et validé dans de nombreuses études. Par exemple, une étude de Revelle et Zinbarg (2009) démontre qu'il s'agit du meilleur indice de fidélité, parmi 12 au total. D'autres études ont depuis confirmé ces résultats, venant confirmer l'omega de McDonald comme le coefficient le plus adéquat pour juger, avec précision, de la fidélité d'échelle de personnalité (Kelley & Pomprasertmanit, 2015 ; Trizano-Hermosilla & Alvarado, 2016; Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019). Il est maintenant souvent recommandé en place de l'alpha de Cronbach.

Depuis, d'autres méthodes ont également vu le jour, afin de contourner les difficultés rencontrées par l'alpha (Green et Yang, 2009; Osburn, 2000; Revelle et Zinbarg, 2009; Sijtsma, 2009; Trizano-Hermosilla et Alvarado, 2016), et notamment les indicateurs lambda2, lambda4 et lambda6 (voir tableau de définition ci-après). Ces mesures sont notamment issues des premiers travaux de Guttman (Guttman, 1945), qui a identifié 6 types de coefficients (lambda 1 à 6), et a montré que chacun était une limite inférieure de la fidélité, définie comme le rapport de la variance du score réel à la variance du score observé (Guttman, 1945; Callender & Osburn, 1979). Comme synthétisé par Bourque, Doucet, LeBlanc, Dupuis et Nadeau (2019) : « *lambda1 sous-estime largement la fidélité réelle et n'est pas utilisé comme estimateur de fidélité, mais comme étape intermédiaire pour d'autres calculs* » (p. 82), (2) lambda3 est équivalent à l'alpha de Cronbach d'un point de vue mathématique, (3) « *lambda-5, quant à lui, est efficace lorsqu'il existe une covariance élevée entre un item et les autres, lesquels, en retour, n'ont pas une covariance élevée entre eux, ce qui n'est pas désirable dans le cas d'une échelle psychométrique* » (p. 83). Parmi ces indicateurs lambda, nous privilégions ainsi ceux qui disposent du plus grand appui empirique quant à leur capacité à estimer la fidélité réelle, à savoir les lambda2, lambda 4 et lambda6. Ceux-ci sont davantage définis dans le tableau ci-après.

Indicateur	Définition	Études
lambda2	Lambda2 est une limite inférieure de fidélité, qui est égale à la vraie fidélité si les composants sont tau-équivalents. Lambda2 est intéressant car il donne toujours une borne inférieure qui est aussi bonne qu'alpha, mais peut être nettement meilleur dans d'autres cas. Le lambda2 est toujours plus élevé que le lambda1 et est supérieur ou égal au lambda3 (donc à l'alpha de Cronbach) en cas d'indépendance entre les erreurs des items.	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 est calculé en divisant le questionnaire en deux moitiés (méthode split-half). Ensuite, la covariance entre les scores obtenus à chaque moitié de questionnaire est calculée. La variance du score total du questionnaire (comprenant les deux moitiés) est également calculée. Lambda4 est généralement considérée en prenant le split qui maxime la fidélité. Le lambda4 représente donc un coefficient de bissection.	
lambda6	Lambda-6 reflète la proportion de variance moyenne de chaque item expliquée en régressant cet item sur tous les autres items, c'est-à-dire la corrélation multiple au carré.	

Table 7.1. Définition des indicateurs lambda.

Il est important de noter que ces indicateurs ne sont pas interchangeables et que leur choix dépendra des objectifs de l'étude et des caractéristiques de l'échelle de mesure. Toutefois, plusieurs recherches ont montré que : (1) l'alpha de Cronbach serait l'un des indicateurs les moins efficaces, (2) l'alpha de Cronbach et le lambda2 sous-estiment systématiquement et significativement la fidélité, (3) le meilleur indice serait l'omega dans le cas où il y a peu d'items, (4) et le lambda6 dans tous les autres cas (Bourque, Doucet, LeBlanc, Dupuis & Nadeau, 2019). Aussi, bien que lambda4 soit un coefficient de fidélité qui reste intéressant en termes de facilité de compréhension, et qu'il est moins susceptible de sous-estimer la fidélité comme peut le faire l'alpha de Cronbach, il peut avoir tendance à surestimer la fidélité s'il y a un grand nombre d'items, ou si la taille de l'échantillon est petit (Benton, 2013; Berge & Socan, 2004) : toutefois, au regard de la nature du questionnaire SWIPE, composé d'un nombre restreint d'items, et de l'échantillon assez large utilisé pour nos études, ce risque est minimisé. En somme, bien que tous ces indicateurs soient proposés dans le cadre des études SWIPE, les plus cohérents sont ici l'omega, le lambda4 et le lambda6.

Par ailleurs, une étude des erreurs de mesures est également proposée (Kim & Feldt, 2010). L'erreur de mesure pour un test de personnalité est la variation aléatoire de la mesure de la personnalité qui peut se

produire en raison d'erreurs de mesure ou de facteurs extérieurs qui affectent les résultats du test. Cette erreur peut être due à plusieurs facteurs, tels que des différences individuelles dans la compréhension des questions du test, des erreurs de notation ou de codage des réponses, des variations dans l'état d'esprit ou l'humeur de la personne testée, ou encore des erreurs dans la méthode de mesure utilisée. L'erreur de mesure peut avoir un impact sur la fidélité et la validité des résultats du test de personnalité, car elle peut conduire à des scores qui ne reflètent pas fidèlement les facettes de personnalité de la personne testée. Il est donc nécessaire de la minimiser. Aussi, pour calculer la fidélité de SWIPE relative à l'erreur de mesure, nous privilégions plusieurs analyses :

- La présentation des courbes d'information et d'erreur de mesure pour chaque facette ;
- Le calcul de l'empirical reliability (empirical_rxx), qui fait référence à la fidélité d'un questionnaire mesurée à partir des données empiriques obtenues lors de l'administration du test à un échantillon de personnes ;
- Le calcul de la marginal reliability (marginal_rxx), qui est une estimation de la fidélité d'un questionnaire calculée à partir d'un modèle statistique qui prend en compte la structure des scores du test et les erreurs de mesure. La marginal reliability est considérée comme une estimation théorique de la fidélité d'un test.

Les seuils d'acceptabilité pour chacun de ces indicateurs ont varié au cours du temps, dépendent du type de questionnaire, du nombre d'items ou encore de la distribution des réponses des participants. Néanmoins, il est généralement recommandé des valeurs de : (1) .6-.7 pour l'alpha de Cronbach (Nunnally, 1978), (2) de .7 pour l'omega de McDonald (McDonald, 1999), (3) de .6 pour les indicateurs lambda (Callender & Osburn, 1979), (4) de .6-.7 pour l'empirical_rxx et le marginal_rxx (Chalmers, 2012).

Enfin, une étude de la saturation inter-items est réalisée. Elle indique dans quelle mesure les différents items qui mesurent une facette de personnalité sont liés les uns aux autres. Plus précisément, la saturation inter-items est la corrélation moyenne entre chaque item. Elle mesure la cohérence interne du questionnaire et permet de déterminer dans quelle mesure les items mesurent la même chose. La saturation inter-items moyenne est considérée comme un indicateur plus simple de la cohérence interne d'une échelle que l'α de Cronbach, au sens qu'elle minimise les effets du nombre total d'items. Généralement, un niveau d'homogénéité dit optimal est atteint quand la saturation inter-items pour une facette est comprise entre .15 et .40 (Piedmont & Hyland, 1993). Toute les valeurs en dessous de .1 laisse penser que les items sont trop différents et mesurent des construits différents. À l'inverse, une saturation inter-items supérieure à .4 montrent que les items sont trop similaires et sont redondants. Globalement, le seuil d'acceptabilité se situe donc entre .15 et .50 (Clark & Watson, 1995).

7.1.1. Alpha de Cronbach

Les dernières études d'alpha de Cronbach pour SWIPE ont été conduites en avril 2023 (N = 4457). À l'exception de 5 dimensions, qui présentent des α compris entre .65 et .70, les coefficients α sont tous supérieurs à .70, montrant des résultats de fidélité adéquats du questionnaire SWIPE. Qui plus est, ces résultats sont à mettre en perspective au regard de la structure courte, à choix forcé et multidimensionnelle de SWIPE, qui fait de l' α de Cronbach un indicateur dont le pertinence n'est ici pas optimale.

Facettes	α	Facettes	α
Assertiveness	.75	Aesthetic sensitivity	.70
Energy level	.77	Creative imagination	.76
Sociability	.75	Intellectual curiosity	.66
Compassion	.70	Organization	.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. Alpha de Cronbach (α) pour chaque facette SWIPE.

7.1.2. Omega de McDonald

Les dernières études de l'omega de McDonald pour SWIPE ont été conduites en avril 2023 (N = 4457). Les coefficients ω sont tous supérieurs à .70, montrant des résultats de fidélité adéquats du questionnaire SWIPE. Aussi, au regard de la pertinence de l'omega de McDonald dans le cadre des analyses de fidélité, ces résultats semblent davantage adaptés pour juger, avec pertinence, la fidélité et de la cohérence interne de SWIPE. Ces conclusions démontrent la cohérence des échelles de SWIPE.

Facettes	ω	Facettes	ω
Assertiveness	.78	Aesthetic sensitivity	.74
Energy level	.81	Creative imagination	.79
Sociability	.79	Intellectual curiosity	.70
Compassion	.72	Organization	.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. Omega de McDonald (ω) pour chaque facette SWIPE.

7.1.3. Indicateurs lambda

Les dernières études relatives aux indicateurs lambda pour SWIPE ont été conduites en avril 2023 (N = 4457) et concernent les lambda2, lambda4 et lambda6. Dans l'ensemble, toutes les valeurs sont supérieures à .70. Les valeurs les plus faibles concernent les facettes trust, sincerity et intellectual curiosity, mais restent très proches de .70. Aussi, ces résultats démontrent la fidélité globale du questionnaire SWIPE.

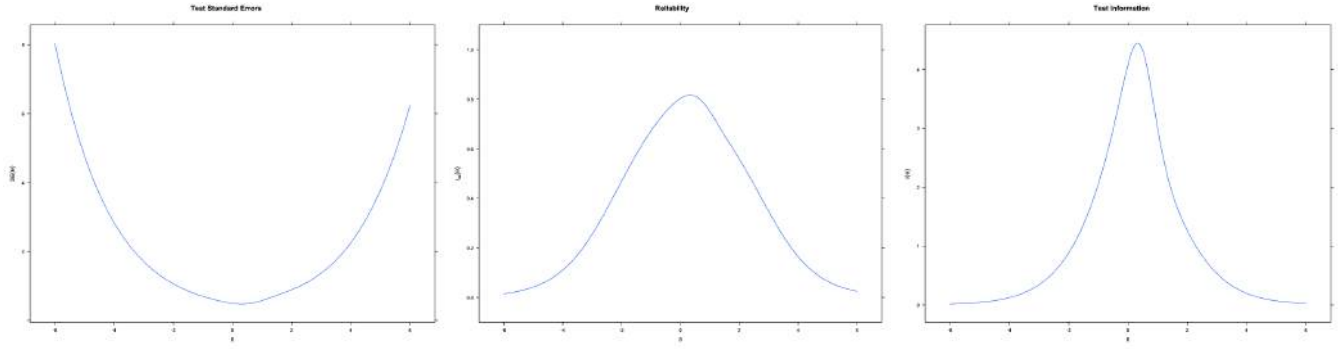
Facettes	lambda2	lambda4	lambda6
Assertiveness	.75	.79	.75
Energy level	.78	.84	.78
Sociability	.76	.80	.75
Compassion	.70	.73	.68
Respectfulness	.68	.72	.67
Trust	.66	.72	.65
Greed avoidance	.70	.74	.68
Modesty	.73	.77	.72
Sincerity	.65	.71	.63
Aesthetic sensitivity	.70	.73	.68
Creative imagination	.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 et lambda6 pour chaque facette SWIPE.

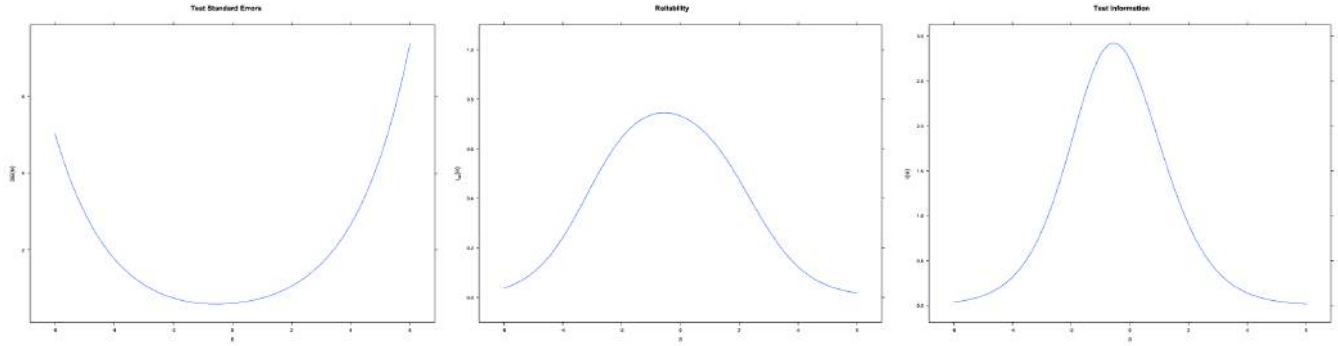
7.1.4. Erreur de mesure et reliability IRT

Les dernières études relatives aux erreurs de mesure, à l'empirical et la marginal reliability pour SWIPE ont été conduites en avril 2023 (N = 4457). Pour chaque facette sont présentés : (1) l'erreur de mesure - test standard errors SE(θ), (2) la reliability rxx(θ) - reliability, (3) la courbe d'information - test information.

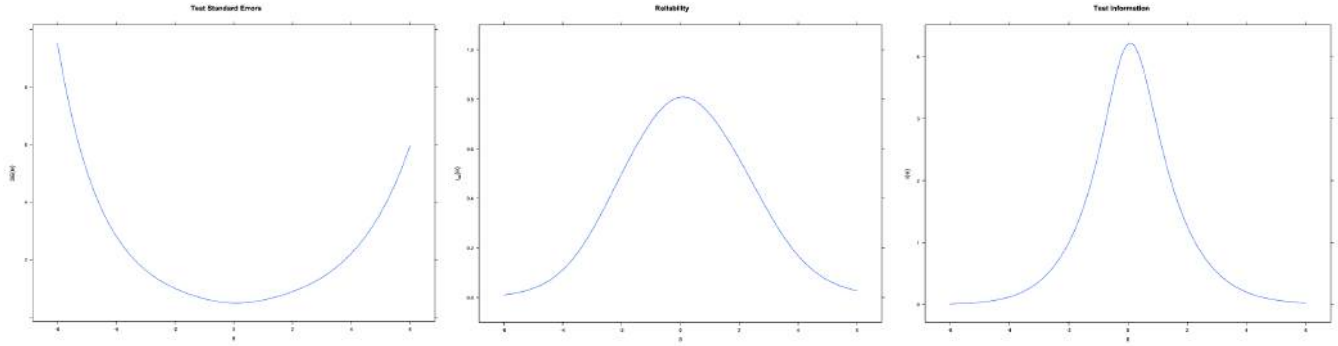
Assertiveness.



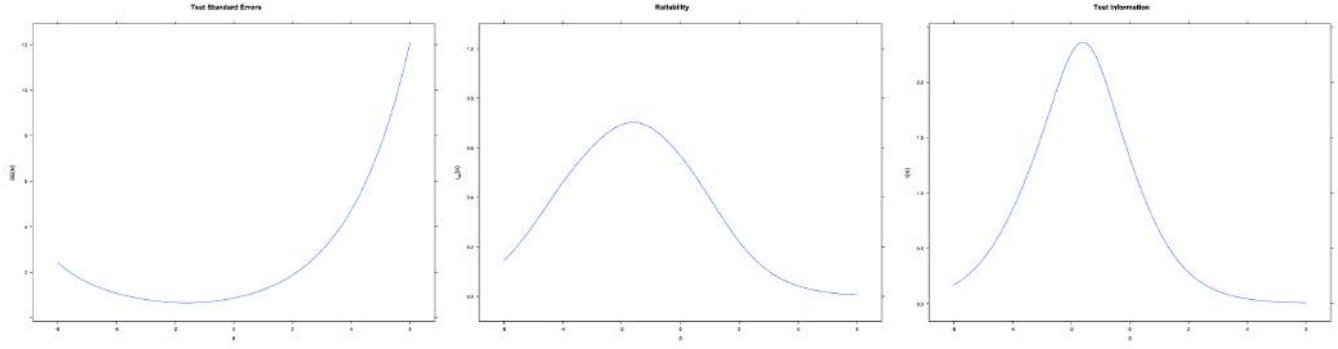
Energy level.



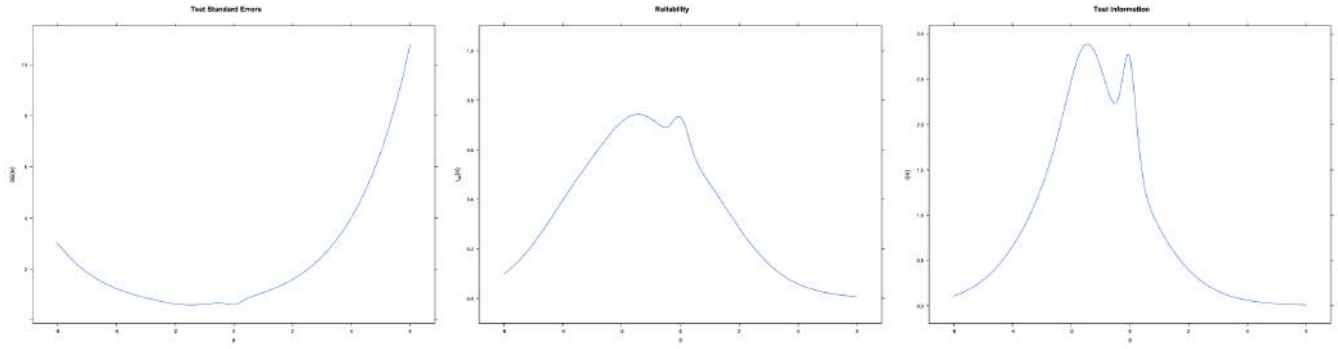
Sociability.



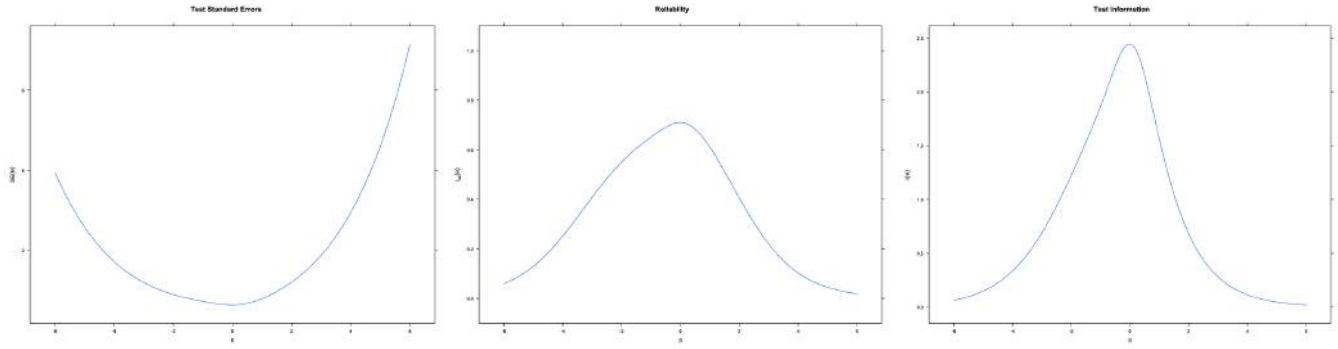
Compassion.



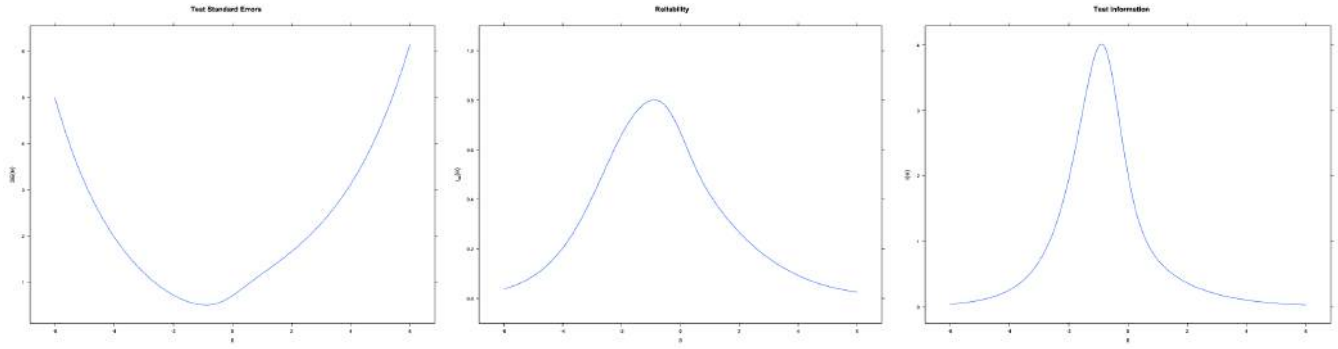
Respectfulness.



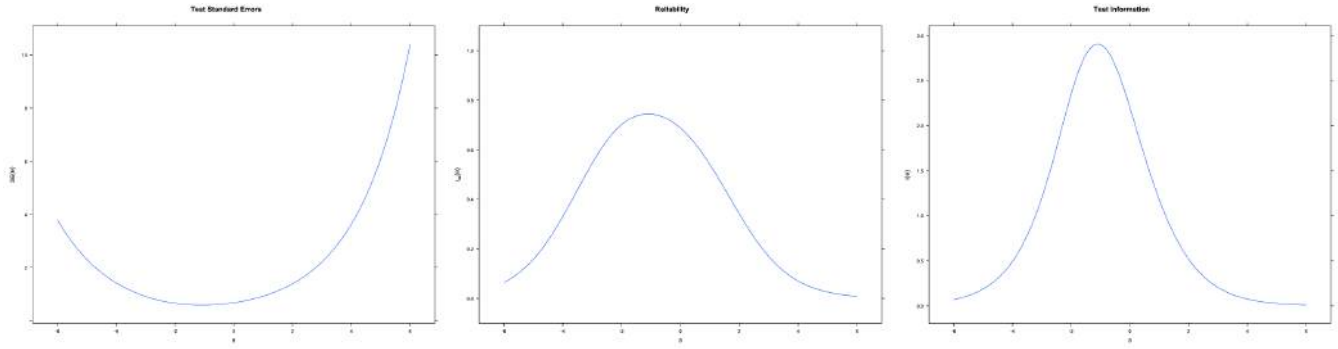
Trust.



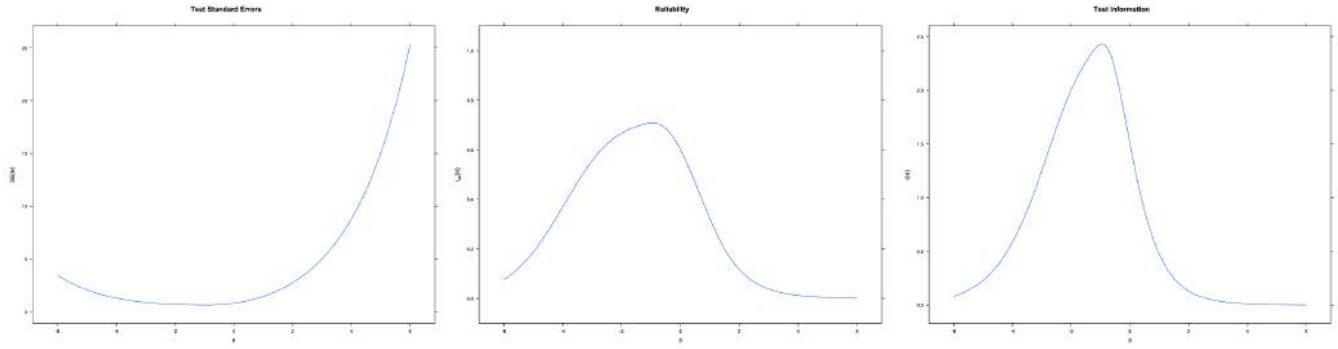
Greed avoidance.



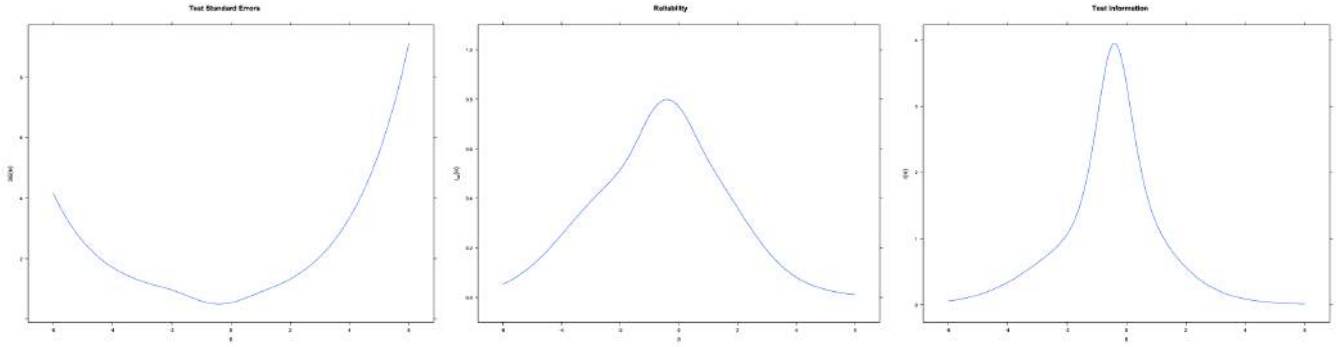
Modesty.



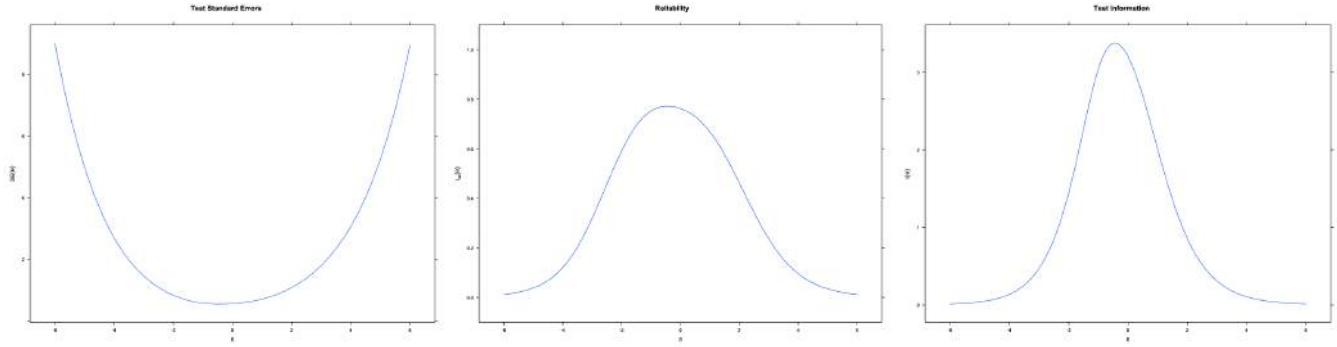
Sincerity.



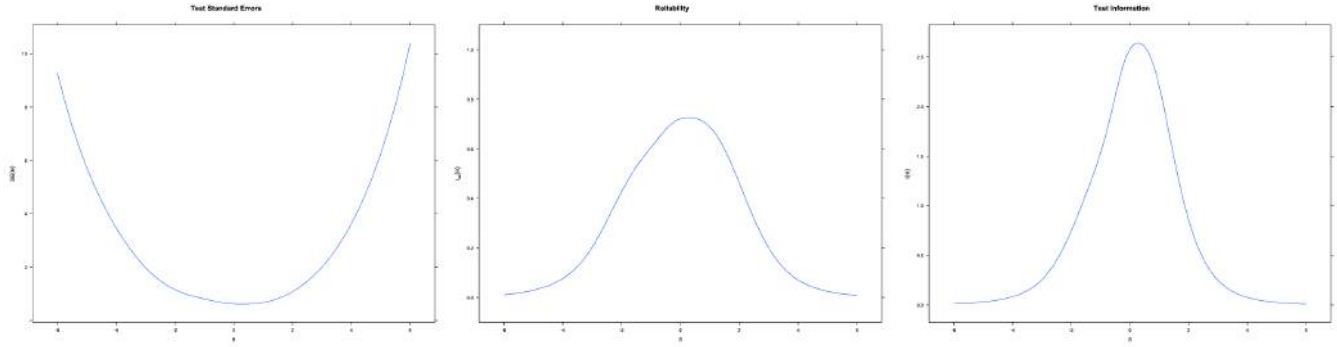
Aesthetic sensitivity.



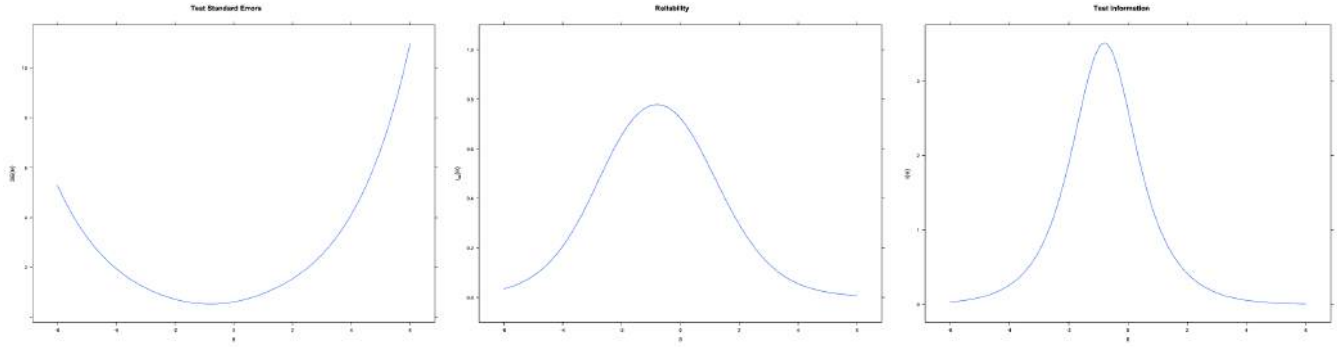
Creative imagination.



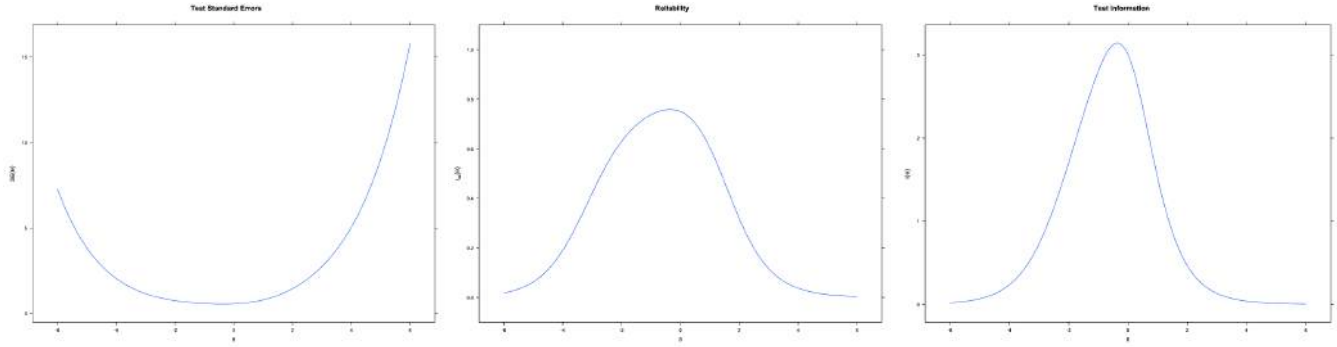
Intellectual curiosity.



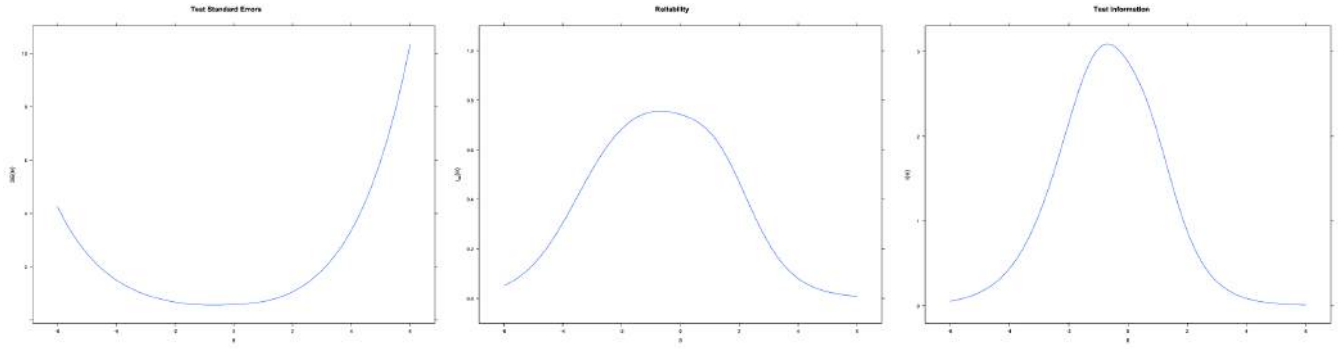
Organization.



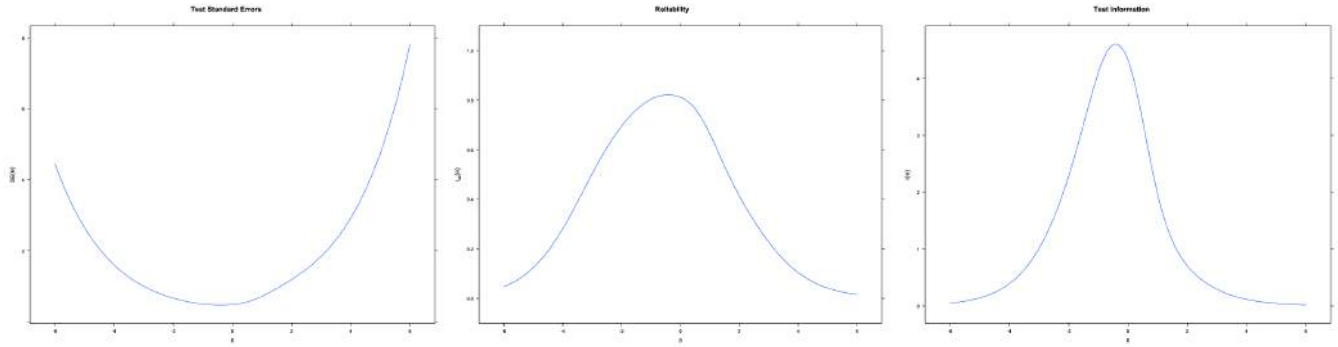
Productiveness.



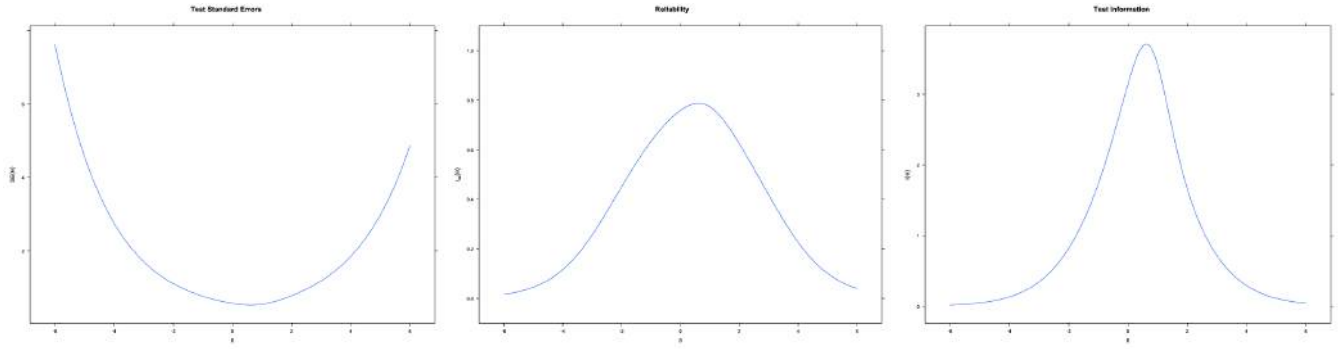
Responsibility.



Anxiety.



Depression.



Emotional stability.

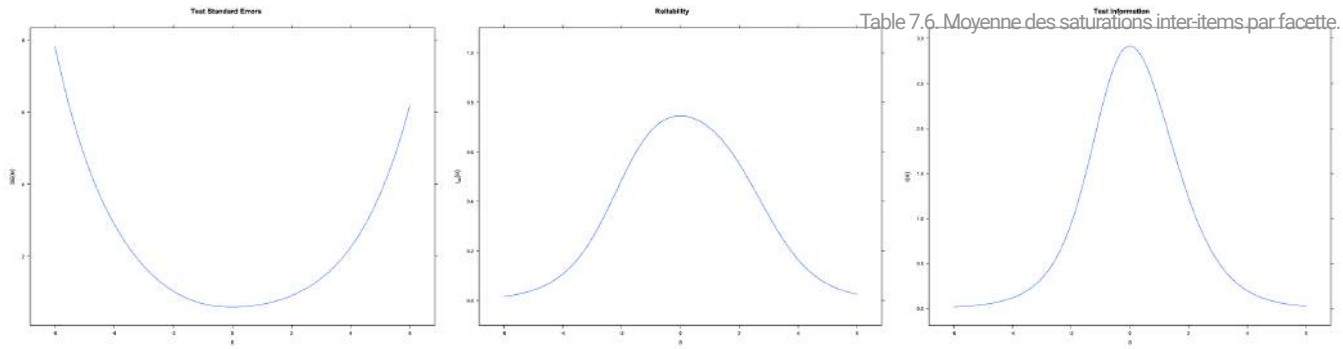


Table 7.6. Moyenne des saturations inter-items par facette.

Dans les tableaux suivants, sont présentés les valeurs de l'empirical_rxx et de la marginal_rxx. Pour ces deux indicateurs, les recommendations de valeurs seuils sont de .6-.7 (Chalmers, 2012). Hormis la facette de personnalité Sincerity (empirical_rxx = .57 et marginal_rxx = .53), l'ensemble des facettes répondent à ces recommandations de seuil, apportant une preuve de fidélité supplémentaire au questionnaire SWIPE.

Facettes	empirical_rxx	marginal_rxx
Assertiveness	.74	.72
Energy level	.73	.70
Sociability	.75	.73
Compassion	.60	.54
Respectfulness	.63	.62
Trust	.65	.64
Greed avoidance	.64	.61
Modesty	.70	.64
Sincerity	.57	.53

Facettes	empirical_rxx	marginal_rxx
Aesthetic sensitivity	.70	.67
Creative imagination	.73	.71
Intellectual curiosity	.66	.65
Organization	.70	.66
Productiveness	.70	.70
Responsibility	.72	.71
Anxiety	.77	.74
Depression	.76	.70
Emotional volatility	.70	.70
	.70	*.67*

Table 7.5. Empirical_rxx et marginal_rxx pour chaque facette SWIPE.

7.1.5. Saturation inter-items

Les dernières études de saturation inter-items pour SWIPE ont été conduites en avril 2023 (N = 4457). Les coefficients MIC sont tous compris entre .15 et .50, ce qui montre un niveau optimal d'homogénéité pour chaque facette du questionnaire (Piedmont & Hyland, 1993; Briggs & Cheek, 1986; Clark & Watson, 1995). En somme, les items mesurant une facette spécifique sont bien liés entre eux, mais pas trop afin d'éviter une redondance. Chaque item apporte ainsi un niveau d'information qui lui est unique et spécifique.

Facettes	MIC	Facettes	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

7.1.6. Conclusion

Les résultats des analyses de cohérence interne présentés indiquent que SWIPE est fiable et répond aux standards exigés en matière de qualité et de fidélité de mesure. Les coefficients alpha de Cronbach et omega de McDonald sont tous élevés, ce qui signifie que les items du questionnaire sont fortement liés les uns aux autres, et mesurent une même dimension de la personnalité de manière consistante. Cette conclusion est d'autant plus intéressante au regard des biais de sous-estimation de l'alpha de Cronbach. Aussi, si les conclusions de l'analyse de l'alpha de Cronbach sont satisfaisantes, celles liées à l'omega de McDonald sont davantage adaptées au contexte de la mesure et des échelles de SWIPE. De même, les indicateurs lambda, qui sont plus adaptés (notamment les lambda4 et lambda6), viennent confirmer et démontrer la cohérence interne du questionnaire SWIPE, et donc sa fidélité globale.

7.2. Fidélité test-retest

La fidélité test-retest est une mesure de la cohérence temporelle d'un questionnaire ou d'une échelle de mesure. Elle consiste à administrer le même questionnaire à un groupe de participants à deux moments différents, avec un intervalle de temps entre les deux administrations. La corrélation entre les deux résultats est ensuite calculée pour déterminer la fidélité du test. Si la corrélation est élevée, cela signifie que les scores des participants sont stables au fil du temps et que le questionnaire peut donc être considéré comme fiable. La fidélité test-retest est particulièrement importante pour les questionnaires de personnalité, car elle permet de s'assurer que les résultats sont cohérents et fiables sur le long terme (Spearman, 1904; Thorndike, 1918; Guilford, 1936; Anastasi, 1954; Cronbach, 1951). Plusieurs études récentes ont d'ailleurs cherché à analyse la fidélité test-retest du BFI et du BFI-2, montrant par exemple une fidélité test-retest élevée pour les cinq domaines de personnalité, avec des corrélations allant de .63 à .86 (Seybert & Becker, 2019; Courtois, Petot, Lignier, Lecocq & Plaisant, 2018; Zhang, Li, Li, Luo, Ye, Yin, Chen, Soto & John, 2022; Courtois, Petot, Plaisant, Allibe, Lignier, Réveillère, Lecocq & John, 2020, Gnambs, 2016).

Les études de fidélité test-retest sont en général menées à des intervalles de 3, 6 et 9 mois. Aussi, SWIPE étant un nouveau questionnaire, les intervalles de temps sont actuellement trop restreints pour pouvoir demander à un échantillon de personnes de repasser le questionnaire. Les études de fidélité test-retest liées à SWIPE seront donc réalisées dans les intervalles de temps opportuns pour pouvoir réaliser une étude fiable : en juillet 2023 (t+3 mois), en octobre 2023 (t+6 mois) et en janvier 2024 (t+9 mois). Les résultats de ces études seront ajoutés à ce manuel technique dès que possible.

Synthèse de fidélité

La fidélité permet de savoir si une mesure est fiable à chaque fois que ce même questionnaire est complété par une même personne. Elle vise à déterminer si un questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne.

.72 Alpha de Cronbach moyen, montrant des résultats de fidélité adéquats du questionnaire SWIPE, malgré la sous-estimation de cet indicateur.

.76 Omega de McDonald moyen, démontrant la forte cohérence des échelles de SWIPE. L'indicateur est plus adapté aux formats multidimensionnels.

.77 Lambda4 moyen. Le lambda4 est davantage adapté à la nature du questionnaire, et permet une nouvelle fois d'attester de la fidélité de SWIPE.

8. Sensibilité

La sensibilité, aussi appelée discrimination, désigne la capacité d'un questionnaire à distinguer les personnes ayant un niveau élevé sur une facette et les personnes ayant un niveau faible. Elle reflète donc la capacité du questionnaire à identifier la singularité de chaque individu. Un questionnaire de personnalité sensible peut ainsi identifier les différences subtiles entre les personnes, et aider à comprendre leurs comportements avec davantage de précision et de discrimination. La sensibilité se réfère ainsi à la capacité du questionnaire à identifier correctement les personnes qui possèdent la caractéristique mesurée et à éviter les faux positifs (personnes qui ne possèdent pas la caractéristique, mais qui sont identifiées comme telles par le questionnaire). Cette propriété est directement liée à la qualité du score.

Les premières tentatives de mesure de la discrimination étaient basées sur des échelles cumulatives (Guttman, 1944; Walker, 1931; Loevinger, 1948; Loevinger, 1953, cités par Hankins, 2008). Ferguson est toutefois l'un des premiers à proposer de conceptualiser la discrimination sous la forme d'un coefficient. En ce sens, s'il existe un nombre maximum de différences possibles dans un échantillon, le coefficient de discrimination correspond au ratio entre le nombre de différences réellement constatées, et ce nombre maximal de différences. Ce coefficient, appelé delta de Ferguson (Ferguson, 1949; Kline, 2000), est ainsi le rapport entre les différences observées entre les personnes et le nombre de différences maximum possibles. Il se veut être un indice direct et non-paramétrique du degré de distinction faite par un instrument entre des individus. Si aucune différence n'est observée, alors $\delta = 0$. Si toutes les discriminations possibles sont observées, alors $\delta = 1$. Généralement, une distribution normale devait avoir une excellente discrimination, où $\delta \geq .9$ (Ferguson, 1949). Les discriminations plus faibles étant attendues pour les distributions leptokurtiques (car ces distributions ne parviennent pas à discriminer autour de la moyenne) et les distributions asymétriques (car celles-ci ne parviennent pas à discriminer à une extrémité de la distribution). Démontrer l'excellente discrimination des échelles d'un questionnaire requiert donc un $\delta \geq .9$ pour chaque échelle de mesure. Par exemple, dans une récente étude d'adaptation du BFI-2 en Russe, Kalugin, Shchebetenko, Mishkevich, Soto et John (2021) ont montré que toutes les échelles possédaient de fortes discriminations.

Les dernières études sensibilité de SWIPE, basées sur le calcul du δ de Ferguson, ont été conduites en avril 2023 (N = 4457). Les coefficients δ sont tous supérieurs à .9, montrant une excellente discrimination des échelles de mesure. En d'autres termes, cela signifie que le questionnaire est capable de détecter avec précision les différences individuelles dans la personnalité des personnes testées, et qu'il est sensible aux variations individuelles dans les facettes de personnalité mesurées. Les résultats sont présentés dans le tableau 8.1.

Facettes	δ	Facettes	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 de Ferguson (δ) pour chaque facette SWIPE.

Comment bien lire les résultats : le δ de Ferguson est le rapport entre les différences observées entre les personnes et le nombre de différences maximum possibles. Un δ de 0 signifie que 0% de toutes les discriminations possibles sont assurées par l'échelle, tandis qu'un δ de 1 signifie de 100% de toutes les discriminations possibles sont faites. Par exemple, pour la facette « Sociability », δ = .96, ce que signifie que 96% de toutes les discriminations possibles sont faites par l'échelle « Sociability ».

9. Équité

L'équité dans le contexte d'un questionnaire de personnalité fait référence à la mesure dans laquelle celui-ci est équitable et impartial pour toutes les personnes qui le passent, quelle que soit leur origine, leur sexe, leur orientation sexuelle, leur race ou leur culture. En d'autres termes, un questionnaire équitable est conçu pour être objectif et juste pour toutes les personnes qui le passent, sans biais ou discrimination à l'encontre d'un groupe particulier. Nos équipes mettent en ce sens tout en œuvre pour veiller au caractère équitable des questionnaires et des analyses prédictives, et pour s'assurer que l'usage de nos algorithmes dans les processus décisionnels n'amènent pas de discriminations via des biais algorithmiques insoupçonnés. Aussi, dans le cadre du développement de nos questionnaires, les études d'équité sont de deux types : (1) celles liées à l'accessibilité du questionnaire, et (2) celles liées à l'équité dans les résultats au questionnaire.

9.1. Accessibilité de SWIPE

L'expérience utilisateur et l'accessibilité de la solution sont des enjeux d'importance et d'intérêt pour AssessFirst. En ce sens, nous avons à cœur de proposer un processus d'évaluation et une interface de restitution qui soient simples à appréhender et compréhensibles. Les efforts que nous déployons font aujourd'hui d'AssessFirst l'acteur incontournable de l'expérience utilisateur : l'expérience que nous proposons est fluide, transparente, et surtout, elle s'adresse à tous, quel que soit l'âge, le métier, le diplôme, la maîtrise des outils numériques, etc. En témoigne par exemple l'évaluation faite par les candidats de la solution, disponible ici dans les Google Reviews. Les actions mises en place par AssessFirst pour assurer et améliorer l'accessibilité du questionnaire SWIPE incluent :

- **L'orientation professionnelle du contenu** : SWIPE et ses résultats ont été spécifiquement développés pour être pertinents dans un but professionnel. Les dimensions évaluées ont été choisies en raison de leur pertinence pour l'efficacité professionnelle. Les conclusions tirées de l'utilisation d'AssessFirst se limitent à ce cadre précis ;
- **Le niveau de langue** : AssessFirst s'appuie sur une équipe « Localization » composée de psychologues et d'experts de la gestion linguistique, afin de proposer des contenus textuels compréhensibles et accessibles par tous, dans toutes les langues (15 langues disponibles à ce jour). En ce sens, nous travaillons avec des traducteurs natifs afin de construire et de valider l'ensemble de nos contenus ;
- **La validation du questionnaire** : comme le démontre ce manuel technique, SWIPE a été développé pour répondre aux standards psychométriques les plus exigeants en termes de validité, fidélité et sensibilité ;
- **Fairness by design** : Nous construisons nos questionnaires à partir de contenus neutres, dans la mesure où ils ne font référence à aucun code culturel ou social. De plus, les textes à lire ont été grandement limités dans SWIPE, avec -65% de texte que SHAPE par exemple. Ce travail de réduction des volumes de textes permet ainsi de favoriser l'accessibilité de SWIPE à des personnes affectées par des troubles de la lecture ;
- **Text to speech** : AssessFirst a développé son propre outil de text-to-speech, ou lecture automatique des questionnaires. Cette fonctionnalité permet ainsi d'accéder à un assistant vocal qui lit les questions, renforçant dès lors l'accessibilité pour les personnes en situation de handicap visuel ;
- **Gestion des contrastes** : AssessFirst met en place des actions permettant de personnaliser les contrastes et l'affiche des contenus web, afin de les rendre plus faciles à lire pour les utilisateurs ayant des déficiences visuelles ;
- **Intégration des utilisateurs** : De nombreux partenariats sont mis en place avec des clients cible qui proposent la solution à des publics qui pourraient avoir des difficultés d'accessibilité à l'outil (e.g. utilisateurs en situation de handicap, éloignés de l'emploi, publics jeunes, publics éloignés du numérique, publics sans expérience professionnelle). Des échanges réguliers avec ces partenaires et les utilisateurs nous permettent d'améliorer sans cesse la solution afin de répondre au mieux à leurs besoins.

Ces actions catalysent l'accessibilité de la solution AssessFirst. C'est d'ailleurs là une obsession de tous chez AssessFirst : faire en sorte que chacun puisse tirer parti de l'ensemble de ses résultats, afin de mieux comprendre ce qui le rend unique et valoriser ses talents. Aujourd'hui, les résultats que nous obtenons chez nos partenaires font d'AssessFirst l'une des entreprises les plus innovantes et les plus inclusives de la HR Tech. Nous avons impacté la trajectoire de vie de plus de 5 000 000 de personnes. 5 000 000 d'individus qui ont eu l'opportunité d'être considérés pour qui ils sont réellement en tant qu'être humain, au-delà de leur parcours académique, professionnel, de leur âge, ou encore de leur genre. C'est dans ce sens que vont les actions ici présentées : elles viennent soutenir la qualité de l'expérience utilisateur pour tous les publics.

9.2. Équité dans les résultats

Les données présentées dans cette section démontrent qu'il n'existe pas de différences majeures ou de taille d'effet forte dans les résultats au questionnaire SWIPE en fonction des variables de genre et d'âge. Note : seules les informations personnelles nécessaires à l'utilisation correcte d'AssessFirst sont demandées à nos utilisateurs. Par exemple, il n'y a aucune mention de l'orientation religieuse, politique ou sexuelle, quel que soit le moment. Quand il s'agit de l'âge, nous demandons la date de naissance afin de nous assurer que cela n'affecte pas la façon dont les questions sont traitées. De même, toutes les variables ci-après analysées n'interviennent à aucun moment dans le calcul des résultats dans la solution.

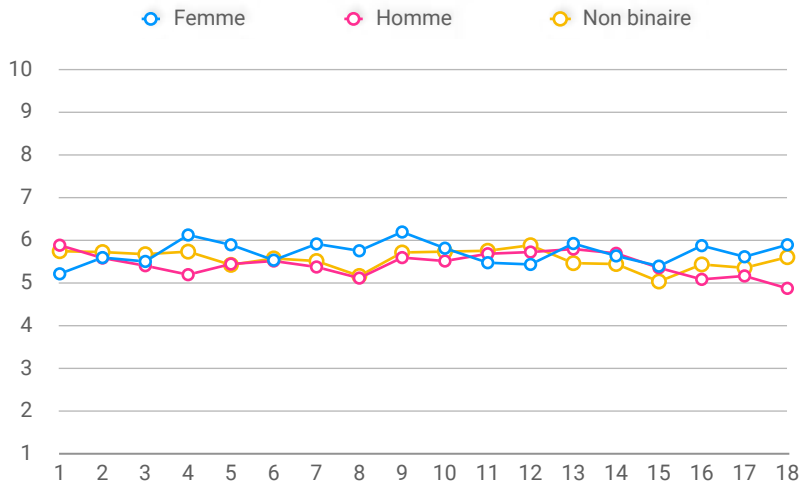
9.2.1. Équité selon le genre

Historiquement, les données relatives à la personnalité sont peu impactées par le genre : en ce sens, hommes et femmes présentent globalement les mêmes comportements. Contrairement à l'opinion populaire, même s'il existe certaines différences, celles-ci sont largement exagérées, et les attributs psychologiques et cognitifs entre genres sont pour la grande majorité similaire : (1) Janet Shibley Hyde (2005), à travers une large méta-analyse pionnière, propose par exemple l'hypothèse de similarité des genres en montrant que pour 78%, les différences entre hommes et femmes sont nulles ou très faibles - notamment quant aux facteurs psychologiques, (2) Ethan Zell, Zlatan Krizan et Sabrina Teeter (2015) renforcent ces conclusions en mettant en avant un chevauchement de 84% dans la distribution des scores entre hommes et femmes, et des effets faibles ou très faibles dans 85% des cas. Aussi, de plus récentes études ont montré que certaines différences entre genre peuvent être le reflet de la manière d'organiser les mêmes données (Eagly & Revelle, 2022) : par exemple, pour les construits psychologiques uniques, les différences entre les genres sont plus prononcées en combinant plusieurs indicateurs qui différent selon le genre pour produire des échelles de typicité globale. Aussi, d'autres études ont quant à elles démontré l'absence d'adverse impact pour les algorithmes basés sur des données liées à la personnalité, avec un impact ratio moyen de .91 (Kubiak, Baron & Niesner, 2023; Efremova, Kubiak, Baron & Frasca, in press).

Ces résultats ne doivent toutefois pas contribuer à cacher de légères différences, naturelles, qui peuvent exister entre les hommes et les femmes sur certaines facettes de personnalité spécifiques : néanmoins, si elles existent, ces différences restent quasi-nulles, négligeables ou faibles. Les femmes ont ainsi tendance à avoir des scores plus élevés que les hommes sur les traits d'agréabilité et de neuroticisme, tandis que les hommes ont tendance à avoir des scores plus élevés que les femmes sur les traits d'extraversion et de conscience. Cependant, les différences de scores entre les sexes sont souvent de petite taille et il y a également une grande variabilité individuelle (Schmitt, Realo, Voracek & Allik, 2008; Weisberg, DeYoung & Hirsh, 2011; Costa, Terracciano & McCrae, 2001; Lippa, 2010, Kajonius & Jonhson, 2018). Aussi, comme le démontrent les indices de taille d'effet de différentes études, ces différences sont modestes et ne concernent généralement que quelques facettes. Par exemple, les effets les plus marqués se retrouvent sur :

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

En somme, les facettes qui semblent plus sensibles aux genres sont les facettes Assertiveness, Anxiety, et Compassion. Au regard de ces constats, il est dès lors probable de retrouver des effets similaires dans le questionnaire SWIPE, avec des tailles semblables. Les dernières études d'équité de genre de SWIPE, basées sur l'analyse des tailles d'effet par le *d* de Cohen, ont été conduites en avril 2023 (N = 3001), sur un échantillon composé de 1624 femmes, 985 hommes et 392 non-binaires. Les résultats sont présentés dans les graphiques (Graph 9.1) et tableaux (Table 9.1). Généralement, une valeur $d \approx .0$ indique qu'il n'y a aucun effet, une valeur $d \approx .3$ correspond à un effet faible, $d \approx .5$ correspond à un effet moyen, et $d \approx .8$ correspond à un effet fort.



Graphique 9.1. Scores moyens aux 18 facettes en fonction du genre.

Globalement, toutes les moyennes se situent entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Comme initialement discutées, les principales différences se trouvent sur Compassion (.93), Anxiety (.78), Assertiveness (.67) et Emotional volatility (1.01). Si ces différences se justifient d'un point de vue théorique, elles sont aussi très certainement dues, pour partie, à l'échantillon utilisé.

Facettes	*d* Cohen	Taille d'effet
Assertiveness	-.35	Faible
Energy level	.01	-
Sociability	.05	-
Compassion	.44	Faible
Respectfulness	.21	Très faible
Trust	.01	-
Greed avoidance	.25	Très faible
Modesty	.34	Faible
Sincerity	.26	Très faible
Aesthetic sensitivity	.14	-
Creative imagination	-.10	-
Intellectual curiosity	-.14	-
Organization	.06	-
Productiveness	-.03	-
Responsibility	.02	-
Anxiety	.40	Faible
Depression	.21	Très faible
Emotional volatility	.54	Moyenne

Table 9.1. *d* de Cohen et taille des effets selon le genre (Homme/Femme).

Les tailles d'effet ici mises en lumière viennent confirmer la littérature sur le sujet : en ce sens, quelques différences existent, notamment sur les facettes liées à l'agréabilité et au neuroticisme. Il convient toutefois de noter que : (1) ces effets sont rares, (2) il s'agit surtout d'effets très faibles ou faibles, (3) ils sont potentiellement inhérents à un effet d'échantillonnage. En effet, l'échantillon faisant ici objet d'analyse à été obtenu suite à sollicitation pour compléter une recherche scientifique en ligne : certaines tendances qui sont spécifiques aux personnalités des personnes qui prennent le temps de répondre à ce genre de démarche ont toutefois été identifiées (Valentino, Zhirkov, Hillygus & Guay, 2020; Marcus & Schütz, 2005). En conclusion, bien que quelques effets soient mis en avant, il n'existe toutefois pas de différence majeure entre les résultats des hommes et des femmes sur les 18 facettes SWIPE. Les résultats de SWIPE sont ainsi équitables selon le genre.

Pour compléter les analyses d'équité de genre, les deux tableaux (voir tables 9.2 et 9.3) suivants présentent les tailles d'effet obtenues suite à (1) la comparaison des femmes et des personnes non-binaires, (2) la comparaison des hommes et des personnes non-binaires. Les effets identifiés sont rares, et ceux identifiés sont très faibles.

Facettes	*d* Cohen	Taille d'effet
Assertiveness	-.28	Très faible
Energy level	-.05	-
Sociability	-.08	-
Compassion	.18	-
Respectfulness	.22	Très faible
Trust	-.02	-
Greed avoidance	.19	-
Modesty	.32	Faible
Sincerity	.21	Très faible
Aesthetic sensitivity	.04	-
Creative imagination	-.14	-
Intellectual curiosity	-.22	Très faible
Organization	.19	-
Productiveness	.09	-
Responsibility	.19	-
Anxiety	.23	Très faible
Depression	.12	-
Emotional volatility	.16	-

Table 9.2. *d* de Cohen et taille des effets selon le genre (Femme/Non-binaire).

Facettes	*d* Cohen	Taille d'effet
Assertiveness	-.07	-
Energy level	-.07	-
Sociability	.13	-
Compassion	-.25	Très faible
Respectfulness	-.00	-
Trust	.03	-
Greed avoidance	.06	-
Modesty	.03	-
Sincerity	.05	-
Aesthetic sensitivity	.10	-
Creative imagination	.03	-
Intellectual curiosity	.08	-
Organization	-.13	-
Productiveness	-.12	-
Responsibility	-.17	-
Anxiety	.17	-
Depression	.08	-
Emotional volatility	.37	Faible

Table 9.3. *d* de Cohen et taille des effets selon le genre (Homme/Non-binaire).

9.2.2. Équité selon l'âge

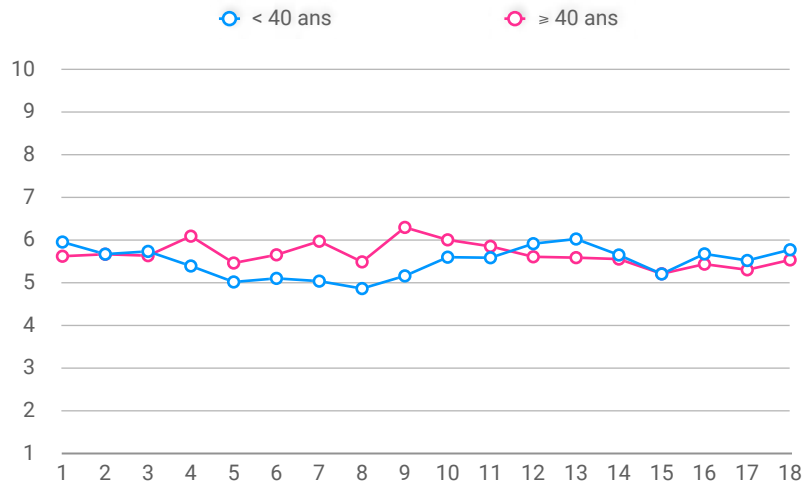
Que sait-on de la stabilité de la personnalité au cours du temps ? Est-ce que la personnalité évolue avec l'âge ? La réponse, issue de plusieurs dizaines d'années d'études en psychologie, est claire : la personnalité n'est pas facilement malléable, mais elle n'est pas non plus forgée dans la pierre. Aussi, selon une méta-analyse d'études longitudinales (Bleidorn, Schwaba, Zheng, Hopwood, Sosa, Roberts & Briley, 2022) :

- Le jeune âge adulte est l'étape de la vie la plus critique pour le développement de la personnalité (Arnett, 2000; Roberts & Mroczek, 2008; Roberts & DelVecchio, 2000; Roberts & Davis, 2016). C'est au début de l'âge adulte que les différences de traits se cristallisent et que la plupart des traits subissent des changements prononcés. Plus précisément, tout au long de l'enfance et de l'adolescence, mais particulièrement lors de la transition vers le jeune âge adulte, les traits deviennent de plus en plus stables avec l'atteinte d'un pic de stabilité vers l'âge de 25 ans ;
- Les estimations de stabilité de personnalité culminent vers l'âge de 25 ans, plafonnent au milieu de l'âge adulte et restent stables ou éventuellement diminuent légèrement à un âge avancé (voir aussi Roberts, Walton & Viechtbauer, 2006; Soto, John, Gosling & Potter, 2011).

En somme, les conclusions de la littérature convergent pour dire que la personnalité devient très stable à partir du jeune âge adulte. Aussi, les changements potentiels à l'âge adulte voient surtout les personnes devenir plus agréables (Roberts, Walton & Viechtbauer, 2006) et plus stables émotionnellement. En

conséquence, les scores moyens aux facettes de personnalité ne devraient pas significativement différer en fonction de l'âge - et ainsi être équitables, bien qu'il pourrait exister une faible tendance à obtenir des scores moyens plus marqués sur les facettes liées à l'agréabilité, et des scores moyens plus faibles sur les facettes liées à la stabilité émotionnelle, pour les tranches plus âgées (Roberts et al., 2006).

Les dernières études d'équité selon l'âge de SWIPE, basées sur l'analyse des tailles d'effet par de d de Cohen, ont été conduites en avril 2023 (N = 306), sur un échantillon avec un âge moyen de 40 ans (σ = 10.97). Sur base de cet âge moyen, deux groupes de comparaison ont été choisis : les personnes dont l'âge est < 40 ans (N = 155) et les personnes dont l'âge est ⩾ à 40 ans (N = 151). Les résultats sont présentés dans les graphiques (Graph 9.2) et tableaux (Table 9.4). Généralement, une valeur $d \approx .0$ indique qu'il n'y a aucun effet, une valeur $d \approx .3$ correspond à un effet faible, $d \approx .5$ correspond à un effet moyen, et $d \approx .8$ correspond à un effet fort.



Graphique 9.2. Scores moyens aux 18 facettes en fonction de l'âge (2 classes).

Globalement, toutes les moyennes se situent entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Les principales différences se trouvent sur des facettes liées à l'agréabilité, et à l'humilité, un domaine proche (Denissen, Soto, Geenen, John & Van Aken, 2022) : Compassion (.69), Sincerity (1.13), Greed avoidance (.93) et Modesty (.62). Celles-ci sont aussi certainement dues à la faible taille de l'échantillon.

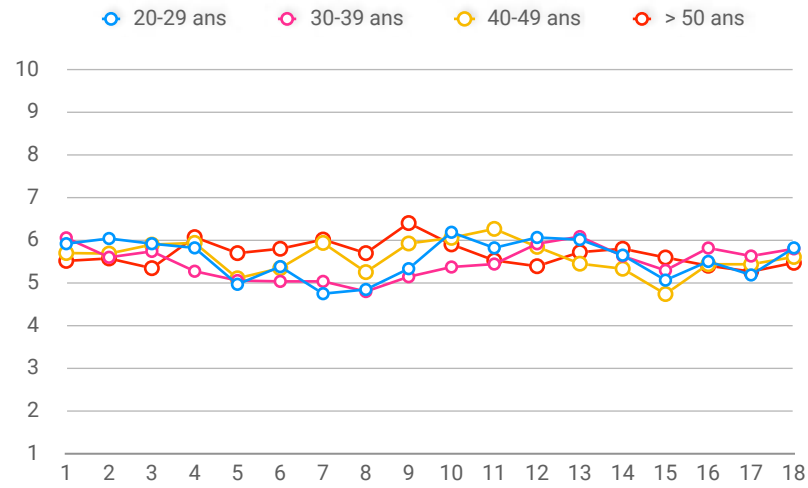
Facettes	d de Cohen	Taille d'effet
Assertiveness	.16	-
Energy level	.00	-
Sociability	.05	-
Compassion	-.30	Faible
Respectfulness	-.19	-
Trust	-.28	Très faible
Greed avoidance	-.46	Faible
Modesty	-.32	Faible
Sincerity	-.49	Faible
Aesthetic sensitivity	-.19	-
Creative imagination	-.11	-
Intellectual curiosity	.14	-
Organization	.18	-
Productiveness	.04	-
Responsibility	-.00	-

Les tailles d'effet ici mises en lumière viennent confirmer la littérature sur le sujet : en ce sens, il n'existe pas de fortes différences selon l'âge. Aussi, même si les effets sont faibles ou très faibles, certaines tendances liées aux facettes d'agréabilité (scores moyens plus hauts pour les ⩾ à 40 ans) et aux facettes de stabilité émotionnelle (scores moyens plus faibles pour les < 40 ans) viennent confirmer les hypothèses initialement proposées (Roberts, Walton & Viechtbauer, 2006). Il convient finalement de noter que : (1) les effets sont rares et concernent uniquement 5 facettes sur les 18 mesurés par SWIPE, (2) il s'agit surtout de tailles d'effets très faibles ou faibles, (3) ils sont potentiellement inhérents à un effet d'échantillonnage, qui est seulement composé de N = 306 personnes. En conclusion, il n'existe pas de différence majeure et significative entre les résultats

Anxiety	.11	-
Depression	.09	-
Emotional volatility	.12	-

Table 9.4. d de Cohen et taille des effets selon l'âge (2 classes).

des ⩾ à 40 ans et des < 40 ans sur les 18 facettes mesurées par SWIPE. Les résultats du sont ainsi jugés équitables selon l'âge.

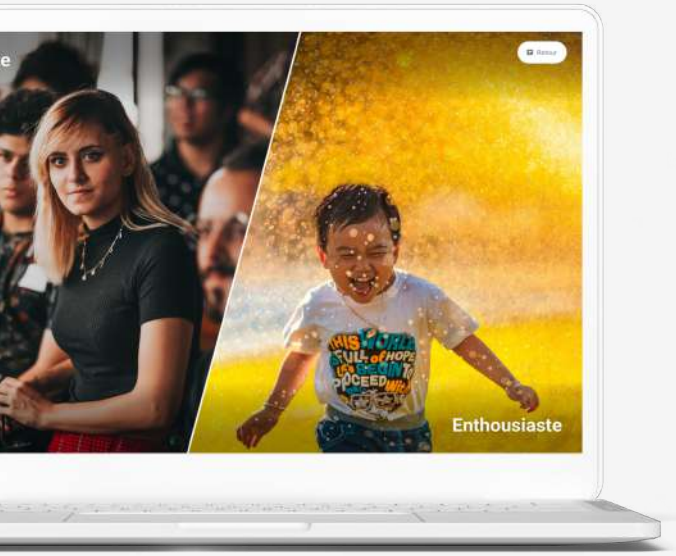


Pour compléter les analyses d'équité d'âge, une analyse plus fine a également été réalisée en prenant en compte 4 catégories d'âges : (1) 20 à 29 ans, (2) 30 à 39 ans, (3) 40 à 49 ans, (4) 50 ans et plus. Globalement, toutes les moyennes se situent entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. De plus, il n'existe pas de différences avec un effet fort et significatif entre ces différentes catégories d'âge.

9.2.3. Conclusion

Que cela soit en termes de genre ou de catégories d'âge, il n'existe pas de différences majeures entre les résultats obtenus par les différents groupes au questionnaire SWIPE d'AssessFirst. En somme, les résultats obtenus permettent de soutenir l'hypothèse selon laquelle le questionnaire proposé ne discrimine aucun public et s'avère équitable. Les effets les plus marqués concernent le genre et uniquement quelques facettes. Aussi, ces effets mentionnés sont très faibles ou faibles, potentiellement explicables par d'autres interactions ou échantillonnages, et trouvent tous une justification conceptuelle.

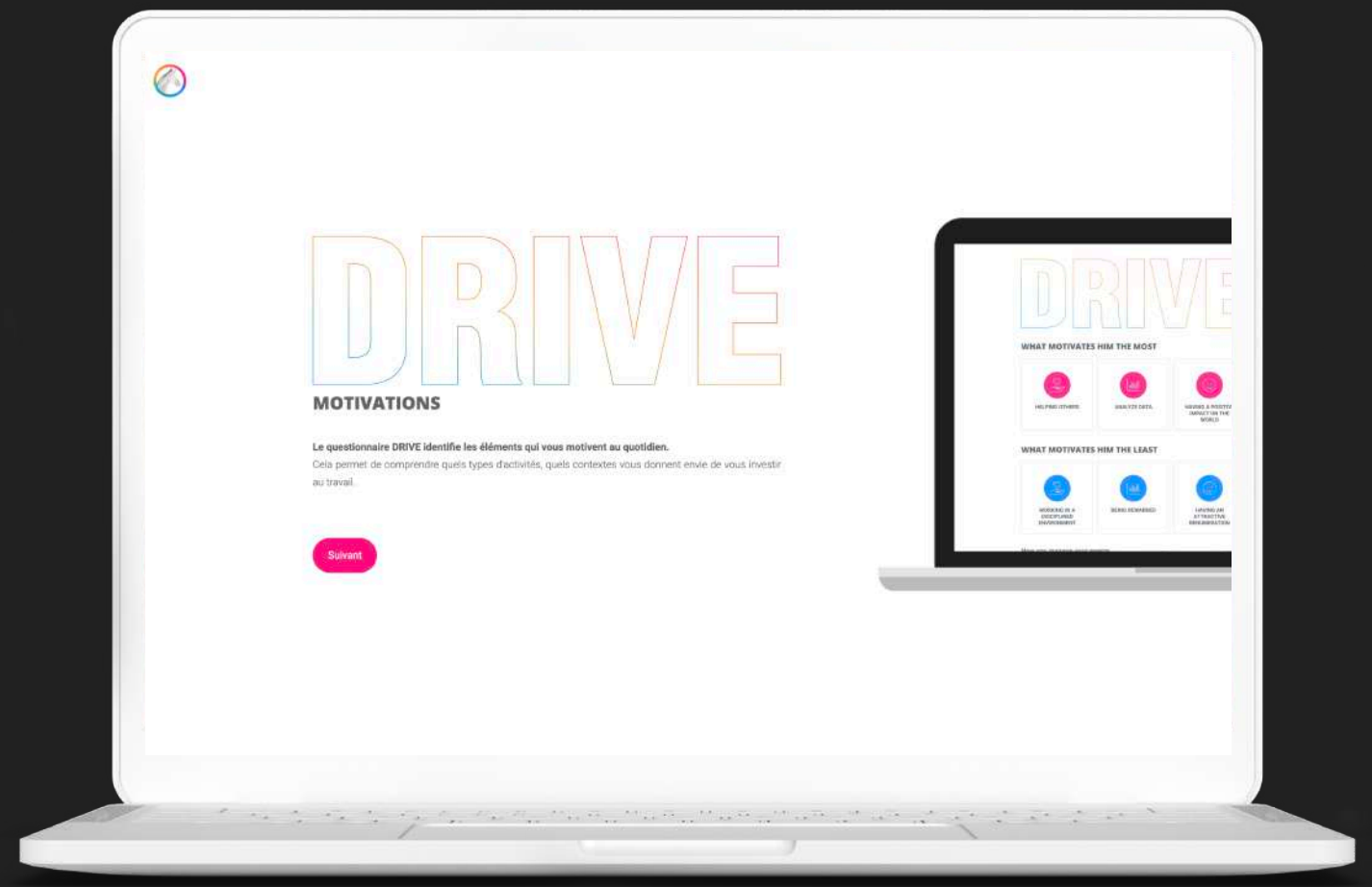
Synthèse d'équité



L'équité dans le contexte d'un questionnaire de personnalité fait référence à la mesure dans laquelle le questionnaire est équitable et impartial pour toutes les personnes qui le passent, quelle que soit leur origine, leur sexe, leur orientation sexuelle, leur race ou leur culture.

.08 *d* de Cohen moyen pour la variable de genre. Les résultats démontrent qu'il n'existe aucun effet du genre sur la plupart des facettes de personnalité.

.08 *d* de Cohen moyen pour la variable d'âge. Les résultats démontrent qu'il n'existe aucun effet de l'âge sur la plupart des facettes de personnalité.



Identifiez les facteurs de motivations

DRIVE est un court questionnaire de motivations, qui permet de comprendre ce qui conditionne la satisfaction et l'engagement des personnes. Réalisable en 10 minutes, DRIVE permet d'anticiper la capacité d'intégration des candidats dans un contexte et une fonction spécifiques.

DRIVE

1. Introduction

DRIVE est un court questionnaire de motivations, qui permet de comprendre ce qui conditionne la satisfaction et l'engagement des personnes. Réalisable en 10 minutes, DRIVE permet d'anticiper la capacité d'intégration des candidats dans un contexte et une fonction spécifiques. Ce questionnaire est composé de **90 items** mesurant **20 dimensions de motivations**. DRIVE a été conçu en 2014 par l'équipe Science d'AssessFirst. Une nouvelle version est en cours de développement pour 2024.

2. Historique de développement

L'évaluation des motivations, qui a émergé comme un champ d'intérêt prédominant en psychologie du travail au siècle dernier, se pose comme question fondamentale : quelles sont les forces qui poussent une personne à se consacrer et à s'engager dans une activité ? Cette interrogation a captivé l'attention de pionniers dans le domaine, qui ont cherché à cartographier le paysage complexe des motivations humaines. Parmi eux, Abraham Maslow a posé les bases avec sa célèbre hiérarchie des besoins en 1954, suivi de près par David McClelland avec sa théorie des besoins en 1961, et Frederick Herzberg avec sa théorie des deux facteurs en 1968. Ces penseurs ont posé les jalons en tentant de formuler des modèles universels des motivations, influençant la compréhension contemporaine de ce qui nous incite à agir dans le monde professionnel.

Les études récentes ont révélé que la motivation est un concept bien plus complexe que ce que les premières théories unidimensionnelles suggéraient. Elle ne saurait être attribuée à un unique facteur, mais résulte plutôt d'un ensemble de facteurs interdépendants. Latham et Pinder, en 2005, ont offert une analyse exhaustive de la littérature existante, aboutissant à des constats significatifs :

- La satisfaction professionnelle est un prédicteur notable de la performance (Judge et al., 2001).
- La motivation au travail réduit la volonté de démissionner (Williams et al., 2001).
- La motivation au travail réduit l'absentéisme et améliore l'engagement (Wegge et al., 2007).

Ces conclusions marquent un tournant dans la compréhension de la dynamique motivationnelle en milieu professionnel et soulignent la nécessité de politiques de gestion des ressources humaines qui reconnaissent la multidimensionnalité de la motivation.

L'ambition d'AssessFirst avec le développement de DRIVE était de capitaliser sur les découvertes académiques concernant la motivation pour en faire un outil pragmatique au service des organisations et des individus. L'idée était de permettre à chacun de prendre des décisions éclairées concernant sa carrière. DRIVE se présente ainsi comme un instrument analytique visant à cerner les facteurs clés qui influent sur la satisfaction et l'implication des employés au travail. Ce faisant, AssessFirst a cherché à offrir aux entreprises un levier pour optimiser le bien-être et la productivité de leurs équipes, en personnalisant l'approche en fonction des motivations de chaque personne.

Le développement de DRIVE par l'équipe Science d'AssessFirst a débuté en 2014, marquant le début d'une phase intensive de recherche et d'innovation. Cette période a été caractérisée par une revue approfondie de la littérature scientifique sur la motivation, l'évaluation des instruments de mesure existants, et une étude détaillée sur les éléments qui définissent l'engagement et le succès professionnel. En début 2016, DRIVE a été lancé sur le marché, offrant ainsi le fruit de ces efforts : un outil conçu pour discerner précisément ce qui propulse les individus vers la réussite dans leurs

fonctions. Cela représente un travail conséquent qui traduit la volonté d'AssessFirst de fournir aux entreprises et aux professionnels des solutions basées sur des données probantes pour favoriser le succès et l'épanouissement dans le milieu professionnel. DRIVE est venu compléter la suite d'outils d'évaluation d'AssessFirst, en s'appuyant sur une décennie et demie d'expertise. Ce nouvel outil a tiré parti des méthodologies éprouvées et des bonnes pratiques développées par l'entreprise, notamment en matière de précision diagnostique, de facilité d'utilisation pour les utilisateurs finaux, et d'interprétation des données et des résultats. De plus, grâce à l'intégration de la théorie de réponse à l'item, un modèle psychométrique de pointe, DRIVE offre des paramétrages avancés pour une évaluation fine et personnalisée des facteurs de motivation et d'engagement au travail. Cela traduit le passage à une approche plus sophistiquée et adaptée aux besoins contemporains des entreprises en matière de gestion des talents.

3. Bases théoriques

La psychologie du travail, au cours des cinquante dernières années, a largement contribué à notre compréhension de la motivation. Chez AssessFirst, nous avons eu le privilège de nous appuyer sur un riche corpus de réflexions théoriques et d'études empiriques pour forger notre questionnaire DRIVE. Parmi l'abondance de recherches, certaines se sont distinguées par leur pertinence et leur proximité avec notre but ultime : décrypter les éléments essentiels qui cultivent la satisfaction et l'engagement des salariés. Ces travaux, qui ont directement influencé notre démarche, nous ont permis de bâtir un outil qui ne se contente pas de mesurer, mais qui éclaire et guide vers une meilleure adéquation professionnelle. Voici un résumé des recherches qui ont été déterminantes dans ce processus.

3.1. Théorie de l’auto-détermination

La conception de DRIVE s'est fortement inspirée de la théorie de l'auto-détermination de Deci et Ryan, qui occupe une place centrale dans le domaine de la psychologie du travail moderne. Cette théorie, largement reconnue pour son autorité académique, soutient que la motivation durable d'un individu ne peut pas être assurée uniquement par des récompenses extrinsèques. Pour qu'une personne reste engagée à long terme dans une activité, elle doit puiser dans ses motivations intrinsèques. En intégrant ces principes, DRIVE cherche à identifier et à évaluer ces motivations intrinsèques, en fournissant un aperçu plus nuancé et profond de ce qui motive réellement les employés au travail, au-delà des simples incitations matérielles. Cela permet aux utilisateurs de DRIVE de comprendre et de cultiver un environnement de travail qui favorise une motivation auto-entretenue, essentielle pour l’engagement.

Selon la théorie de l'auto-détermination, la motivation intrinsèque est alimentée par la satisfaction de trois besoins psychologiques fondamentaux. Premièrement, le besoin de compétence, qui réfère à la capacité d'une personne à réussir et à maîtriser les tâches qui lui sont confiées. Deuxièmement, le besoin d'autonomie, qui souligne l'importance de la liberté de choix et la capacité de l'individu à surmonter les obstacles en faisant preuve d'initiative et de créativité. Enfin, le besoin d'affiliation, qui met en lumière le désir d'être en relation avec d'autres, de se sentir appartenant à un groupe et soutenu par ses pairs. DRIVE intègre ces besoins dans son analyse pour fournir un cadre permettant aux individus et aux organisations de comprendre et de renforcer les motivations intrinsèques.

DRIVE reflète la théorie de l'auto-détermination dans sa structure même. Sur les vingt dimensions évaluées par le test, dix-huit correspondent à des motivations intrinsèques, reflétant la conviction que ces facteurs sont prédominants pour le maintien de l'engagement à long terme. Cependant, conscient de l'importance de ne pas exclure le rôle des motivations extrinsèques, DRIVE inclut également deux dimensions extrinsèques. Cette décision s'appuie sur la méta-analyse de Cerasoli, Nickin & Ford en 2014, qui a démontré que, dans certaines circonstances professionnelles, ces facteurs extrinsèques peuvent

exercer une influence positive. DRIVE est conçu pour évaluer la satisfaction des trois besoins fondamentaux qui sont au cœur de la théorie de l'auto-détermination, assurant ainsi que l'outil est ancré dans une compréhension profonde de ce qui stimule les employés dans leur milieu de travail.

Bien que la théorie de l'auto-détermination fournit un cadre robuste pour comprendre et évaluer la motivation, il est à noter qu'elle n'est pas spécialement conçue pour identifier tous les facteurs spécifiques qui favorisent l'épanouissement individuel au-delà du contexte professionnel. En d'autres termes, cette théorie se concentre sur l'audit du niveau de motivation présente en lien avec l'autonomie, la compétence et l'affiliation, mais ne prétend pas explorer l'ensemble des conditions d'épanouissement personnel qui peuvent être affectés par des facteurs divers et variés non couverts par ce modèle.

3.2. Motives, Values, Preferences Inventory

Le développement de DRIVE a été grandement influencé par l'inventaire de Joyce et Robert Hogan, qui constitue une référence dans le domaine de la psychologie du travail. Leur analyse rigoureuse s'étend sur huit décennies de recherche, ce qui a abouti à un modèle dont l'excellence est reconnue depuis près de vingt ans. Ce modèle se démarque par sa capacité à intégrer de manière transversale les différentes théories et études expérimentales existantes, tout en étant particulièrement adapté aux contextes professionnels, domaine pour lequel il a été spécifiquement conçu. C'est pourquoi les dix dimensions du Motives, Values, Preferences Inventory (MVPI) se retrouvent, de manière indirecte, intégrées dans les fondements de DRIVE, enrichissant ainsi l'outil avec un spectre large de facteurs motivationnels.

3.3. Multidimensional theory of person-environment fit

L'exploration méticuleuse des avancées récentes en psychologie du travail a permis à notre équipe de pousser les frontières de la compréhension traditionnelle de la motivation. Cette investigation nous a guidés vers une conceptualisation plus sophistiquée, en intégrant la modélisation de divers niveaux d'adéquation ou de "fit" entre un individu et son milieu professionnel. Ce concept multidimensionnel de "fit" est central dans DRIVE, car il ne se limite pas à évaluer la motivation en tant que telle, mais cherche également à déterminer comment l'alignement entre les aspirations personnelles, les compétences, les valeurs et l'environnement de travail peut optimiser l'engagement et la satisfaction au travail.

La majeure partie des recherches sur la satisfaction et l'engagement au travail font état de l'impact du lien entre une personne et son environnement de travail. Les recherches significatives de Edwards et Billsberry, publiées en 2010, ont mis en lumière l'importance des différents niveaux de "fit" dans un contexte professionnel. Ces niveaux de "fit", qui s'inspirent du modèle élaboré par Jansen et Kristof-Brown en 2006, comprennent :

- '*Person-vocation fit*' : Cela se réfère à l'adéquation entre les intérêts, les valeurs, les compétences et les aptitudes d'un individu et les exigences ou les récompenses d'une profession ou d'une carrière. Une bonne adéquation ici signifie que la personne est bien adaptée aux tâches, aux valeurs et aux attentes de la profession qu'elle a choisie.

- '*Person-organization fit*' : Ce concept se focalise sur la compatibilité entre les valeurs d'un individu et la culture, les normes et les valeurs d'une organisation. Un bon "fit" signifie que la personne partage des valeurs similaires à celles de l'organisation, ce qui peut entraîner une meilleure satisfaction au travail et une plus grande probabilité de réussite au sein de l'entreprise.

- '*Person-group fit*' : Il s'agit de la correspondance entre un individu et un groupe de travail spécifique au sein de l'organisation, y compris des équipes ou des départements. Une bonne adéquation peut conduire à une meilleure collaboration et communication, et à un meilleur travail d'équipe.

- '*Person-job fit*' : Ce terme décrit la relation entre les compétences, les expériences et les attentes d'un individu et les exigences du poste qu'il occupe. Une bonne adéquation peut se traduire par de meilleures performances, une plus grande satisfaction au travail et une réduction du turnover.
- '*Person-person fit*' : Cela se réfère à la relation interpersonnelle et à la compatibilité entre individus au travail, que ce soit avec des collègues, des supérieurs ou des subordonnés. Un bon alignement peut améliorer la communication, le respect mutuel et le soutien au sein de l'organisation.

Cette compréhension nuancée est incorporée dans DRIVE, permettant ainsi de mesurer et d'analyser ces diverses formes d'alignement pour mieux comprendre et promouvoir l'engagement et la satisfaction de chacun dans son environnement et dans son travail. Les cinq niveaux de "fit" identifiés dans les recherches influencent de manière significative trois indicateurs clés dans le milieu professionnel : l'engagement des employés, leur intention de rester à leur poste ou de le quitter, et leur satisfaction globale au travail. Avec l'objectif affirmé de renforcer ces indicateurs, DRIVE a été élaboré en s'appuyant sur ces modèles éprouvés. Cette stratégie a guidé notre démarche pour créer un outil capable d'adresser et d'améliorer de façon ciblée l'adéquation entre les employés et leur environnement de travail, dans le but ultime de favoriser un milieu professionnel où l'engagement et la satisfaction sont optimisés.

3.4. Les dimensions de DRIVE

Facettes	Définition
Créer de nouvelles choses	Besoin d'exprimer de la créativité dans son travail
Se dépasser au quotidien	Besoin d'être confronté(e) à des défis ambitieux
Se préoccuper de l'esthétique	Besoin de soigner la présentation de son travail
Analyser des données	Besoin de travailler sur des éléments factuels et réaliser des analyses
Rencontrer de nouvelles personnes	Besoin d'avoir un travail centré sur les personnes
Avoir des tâches clairement définies	Besoin d'avancer sur un travail à court terme
Se préoccuper de la qualité	Besoin de réaliser un travail fiable, précis
Avoir de l'influence	Besoin d'avoir du pouvoir dans les décisions
Avoir de l'autonomie	Besoin de liberté d'action
Travailler en équipe	Besoin de réaliser un travail en collaboration avec d'autres
Avoir un impact positif sur le monde	Besoin de s'engager dans une entreprise qui a un effet positif sur le monde
Travailler dans une ambiance fun	Besoin de stimulations sociales, d'un cadre dé travail détendu
Evoluer dans un environnement sécurisant	Besoin d'un environnement stable, un emploi sur le long terme
Travailler de façon disciplinée	Besoin de se référer à des règles et des principes établis
Préserver son équilibre personnel	Besoin d'avoir du temps pour soi, pas que pour le travail
Recevoir des récompenses	Boosté(e) par l'idée d'avoir une récompense à la clé
Avoir une rémunération attractive	Besoin de gagner beaucoup d'argent, amasser de la valeur
Remporter régulièrement des succès	Besoin de victoires, de gagner, de réussir dans les activités
Aider les autres	Besoin d'apporter son soutien aux autres
Être reconnu(e) par les autres	Besoin d'être valorisé(e) par les autres

Table 3.1. Dimensions de motivations mesurées par DRIVE.

4. Construction de DRIVE

La conception de DRIVE s'est articulée autour de quatre piliers fondamentaux :

- **Ancrage scientifique** : Chaque facteur évalué par DRIVE repose sur des bases scientifiques solides, étayées par des publications de recherche concluantes.
- **Pertinence professionnelle** : Les facteurs mesurés sont en résonance directe avec les réalités du monde du travail actuel et doivent être immédiatement applicables dans des contextes professionnels.
- **Accessibilité du questionnaire** : Le processus d'évaluation est conçu pour être intuitif et fluide, garantissant ainsi que le questionnaire puisse être complété sans difficulté.
- **Clarté des résultats** : Les données issues de DRIVE sont présentées de manière à ce qu'elles soient immédiatement compréhensibles et exploitables par les utilisateurs.

Ces principes ont été essentiels pour élaborer un instrument de mesure qui répond précisément aux exigences et aux attentes des entreprises tout en étant universellement applicable, quel que soit le profil.

4.1. Choix des dimensions

Le développement de DRIVE s'est nourrie d'une exploration exhaustive des connaissances disponibles, incluant des publications scientifiques, des modèles d'entreprise, des outils d'évaluation existants, et des études sur la réussite professionnelle. En s'appuyant sur les principes fondamentaux préalablement définis, l'équipe a minutieusement sélectionné 25 facteurs de motivation qui répondaient à ces exigences rigoureuses. Ces facteurs ont été choisis en raison de leur solidité à la fois dans la littérature scientifique et dans la pratique empirique, garantissant ainsi que DRIVE repose sur un socle de données probantes et d'applications concrètes dans le monde professionnel.

En examinant les outils existants de mesure des motivations, notre équipe a relevé que la plupart se contentent d'analyser un spectre limité, souvent moins d'une dizaine de critères. Bien que cela puisse suffire pour établir une base de la motivation, cela s'avère insuffisant pour déceler les nuances des besoins motivationnels individuels. De manière analogue aux Big Five en psychologie de la personnalité, qui fournissent une structure robuste pour comprendre les traits généraux d'un individu, ces grands domaines nécessitent une analyse plus fine pour obtenir une compréhension approfondie des caractéristiques personnelles. DRIVE est donc conçu pour aller au-delà de cette évaluation de surface, en proposant une exploration détaillée et nuancée des facteurs de motivation. D'autre part, notre analyse a révélé l'existence d'outils extrêmement détaillés, évaluant plus de 30 critères. Ces instruments offrent une compréhension exhaustive des moteurs de la motivation, mais leur complexité rend l'exploitation des résultats difficile et chronophage. Cette réalité est en opposition directe avec l'un de nos principes fondamentaux : fournir des informations claires et aisément exploitables. C'est pourquoi DRIVE a été pensé pour équilibrer la profondeur de l'analyse avec la simplicité d'utilisation, assurant ainsi que les utilisateurs bénéficient d'insights pertinents sans être submergés par un excès de complexité.

La sélection des 25 facteurs initiaux pour DRIVE représente un équilibre judicieux entre la simplification excessive et une complexité superflue. Nous avons méticuleusement élaboré des définitions claires et distinctes pour chacun de ces facteurs, afin d'assurer qu'ils incarnent des concepts précis et autonomes. Cette démarche a permis de créer un cadre d'évaluation qui est à la fois riche en contenu et pratique pour les utilisateurs, évitant ainsi les pièges d'un outil trop élémentaire ou, trop laborieux à interpréter.

4.2. Rédaction des items

Sur la fondation solide que constituent les facteurs et leurs définitions, une équipe de cinq psychologues du travail a procédé à la formulation indépendante des items de DRIVE. Trois directives clés ont orienté leur rédaction pour garantir l'accessibilité et la pertinence de l'outil :

- **Clarté linguistique** : Chaque item devait être rédigé dans un langage simple et clair, afin d'être compréhensible par toute personne, quelle que soit sa familiarité avec les termes psychologiques.
- **Précision sémantique** : Il était impératif d'utiliser des termes non équivoques pour éviter toute ambiguïté qui pourrait compromettre la fidélité des réponses.
- **Pertinence contemporaine** : Les items devaient refléter des situations actuelles et réelles dans l'environnement de travail, pour assurer que l'outil soit ancré dans la réalité professionnelle moderne.

Ces consignes ont permis de créer des items qui non seulement mesurent les motivations de manière précise mais sont aussi intuitifs et directement applicables pour les utilisateurs finaux, renforçant ainsi la valeur pratique de DRIVE. Aussi, dans la rédaction des items pour DRIVE, une stratégie de dualité a été adoptée, inspirée des constatations de Brown & Maydeu-Olivares en 2012. Pour chaque concept mesuré, des items ont été formulés de manière à capturer l'essence du facteur (items positifs) ainsi que des items conçus pour représenter l'inverse ou l'antithèse du même concept (items négatifs). Cette approche binaire a pour but d'enrichir la dimensionnalité de l'évaluation et de minimiser les biais de réponse. En proposant cette alternance entre affirmations positives et négatives, DRIVE accroît la précision et la fidélité de l'évaluation, permettant ainsi une mesure plus équilibrée des motivations.

Pour DRIVE, un processus exhaustif de développement d'items a été mis en place, avec la création d'environ 750 items, répartis sur les 25 facteurs identifiés, ce qui représente approximativement 30 items par facteur. Afin de valider leur qualité et leur pertinence, ces items ont été minutieusement examinés par un groupe externe de 20 observateurs qui n'étaient pas impliqués dans la phase de création. Ces observateurs ont eu pour mission de relier chaque item à un facteur spécifique, garantissant ainsi une association claire et univoque entre l'item et le concept qu'il est censé mesurer. De plus, afin de s'assurer de la clarté de formulation, chaque item a été évalué sur une échelle de 1 (peu clair) à 5 (très clair). Ce processus rigoureux a pour but d'éliminer tout flou et d'assurer que chaque item contribue de manière significative à la mesure précise du facteur de motivation correspondant. La sélection finale des items pour la première version du questionnaire DRIVE s'est basée sur un critère combiné de pertinence et de clarté. Seuls les items qui ont été associés de manière cohérente à leur facteur respectif par au moins 80 % des observateurs, et qui ont obtenu une note moyenne supérieure à 4 sur 5 en termes de clarté, ont été retenus. Cette méthodologie rigoureuse garantit que chaque item inclus dans DRIVE est à la fois précisément lié au concept qu'il est destiné à mesurer et formulé de manière à être aisément compréhensible, assurant ainsi la fidélité et l'efficacité du questionnaire dès son lancement.

4.3. Étapes de développement

Initialement, pour valider les items sélectionnés, une version normative du questionnaire DRIVE a été développée. Celle-ci a été mise à disposition en ligne, permettant aux répondants de se positionner sur une échelle allant de 1 (pas du tout motivant) à 5 (très motivant) pour chaque item. Pour éviter l'épuisement des participants, qui aurait pu introduire un biais dans les résultats, le questionnaire a été scindé en deux versions distinctes, chacune contenant 280 items. Cette division visait à alléger la charge pour les répondants et à garantir l'intégrité des données collectées, car passer un questionnaire de 560 items en une seule fois aurait pu impacter la qualité des réponses du fait de la fatigue.

Chaque version du questionnaire DRIVE, contenant 280 items, a été soumise à un échantillon d'environ 300 individus volontaires. Ces participants, représentant la population française active âgée de 25 à 55 ans et travaillant à temps plein, ont fourni les données initiales pour une analyse préliminaire. Les réponses recueillies ont permis de conduire une étude de validité structurelle, dont les résultats ont été consignés dans la section dédiée à la validité du questionnaire. Cette étude visait à confirmer l'organisation et la cohérence des facteurs de motivation tels qu'ils sont mesurés par DRIVE, assurant ainsi que l'outil respecte les standards psychométriques et répond aux exigences de validité scientifique.

À la suite de l'analyse statistique initiale, il a été relevé que certaines dimensions du questionnaire DRIVE n'offraient pas une fidélité statistique suffisante, ce pour deux raisons principales :

- **Manque de cohérence interne :** Pour quelques facteurs, même avec un nombre conséquent d'items destinés à décrire le concept, il n'était pas possible de distinguer un ensemble d'items cohérents qui seraient fortement corrélés entre eux, ce qui est nécessaire pour évaluer de manière fiable un concept unique.
- **Redondance conceptuelle** : Il a été observé que certains facteurs étaient conceptuellement trop similaires ou opposés, rendant difficile leur distinction en tant qu'entités indépendantes. En conséquence, des décisions ont été prises pour éliminer ou fusionner certains facteurs en consolidant leurs items les plus pertinents et en supprimant les moins discriminants.

Ces ajustements sont essentiels pour garantir que chaque dimension de DRIVE évalue un aspect distinct et pertinent de la motivation, augmentant ainsi la validité et l'utilité du questionnaire pour des applications pratiques dans le milieu professionnel.

Après révision, le questionnaire DRIVE a été affiné à 20 facteurs, avec un total de 372 items soigneusement sélectionnés. Pour améliorer l'accessibilité et la simplicité du questionnaire, une technique de "choix forcé" a été employée, appariant les items par deux, selon leur désirabilité sociale. Cette méthode oblige les répondants à choisir entre deux propositions, ce qui réduit le biais de désirabilité sociale et simplifie le processus de réponse. Cette approche est fondée sur l'expérience et les bonnes pratiques en psychométrie, et a été choisie parce qu'elle s'avère être la plus efficace pour faciliter la passation tout en préservant la qualité des données recueillies. La version finale de DRIVE, structurée autour du format "choix forcé", comportait 186 duos d'items. Cette version a été soumise en ligne à un panel de 347 individus, tous des professionnels français âgés de 25 à 55 ans, travaillant à temps plein. L'analyse des réponses a permis d'évaluer l'équilibre entre les items de chaque paire. L'objectif était d'atteindre une répartition des réponses proche de la parité, avec idéalement 50 % des répondants sélectionnant chacune des deux propositions. Un écart de distribution était toutefois acceptable, avec une tolérance fixée à 60 % pour une proposition et 40 % pour l'autre. Cette démarche assure que chaque couple d'items est équilibré, évitant ainsi un biais potentiel vers une option plus socialement désirable, et contribuant à la précision des mesures de DRIVE.

Pour les items qui ne répondaient pas à la distribution cible de 50/50 ou dans la plage acceptable de 60/40, une révision a été effectuée pour ajuster leur niveau d'attractivité, dans le but de corriger l'asymétrie observée. Si la révision menaçait de compromettre la validité de l'item d'origine, celui-ci était alors écarté pour préserver la fidélité de l'instrument. Cette démarche de raffinement a abouti à une version révisée du questionnaire DRIVE, comprenant 166 paires d'items. Cette nouvelle version a été administrée à un groupe de 341 personnes, sélectionnées pour refléter les mêmes caractéristiques démographiques que les panels antérieurs. Cet ajustement méthodique visait à perfectionner la mesure et à garantir que le questionnaire reste à la fois précis et accessible aux répondants.

Dans la version définitive du questionnaire DRIVE, seuls les 90 couples de propositions offrant les meilleures qualités psychométriques ont été retenus. Cette sélection rigoureuse avait pour but de concentrer l'évaluation sur les items les plus performants et fiables tout en assurant un temps de passation réduit. Cette stratégie répond à l'engagement d'offrir un outil à la fois robuste et respectueux du temps des répondants, facilitant ainsi son intégration dans des contextes professionnels où le temps est un facteur crucial. En résulte un questionnaire optimisé qui capture efficacement les facteurs de motivation sans imposer une charge excessive aux participants.

4.4. Le format du questionnaire

L'adoption d'un format de questionnaire à choix forcé à deux propositions pour DRIVE s'est basée sur les avantages significatifs que ce type d'approche présente :

- **Réduction de la désirabilité sociale** : La structure à choix forcé est reconnue pour son efficacité à minimiser le biais de désirabilité sociale. Dans un contexte de sélection où les répondants pourraient être tentés de se présenter sous leur meilleur jour, il est crucial de le maitriser (Christiansen et al., 2005; Jackson et al., 2000; Martin et al., 2001; Vasilopoulos et al., 2006).
- **Facilité et rapidité de passation** : Comparé à une version normative, le format à choix forcé permet un gain de temps. Les observations ont montré qu'un questionnaire à choix forcé peut être complété 30 % plus rapidement, ce qui le rend plus pratique tant pour l'organisation que pour le candidat.
- **Élimination des biais de réponse** : Les formats normatifs peuvent souvent conduire à des biais de centralité (tendance à choisir des options neutres) et d'extrémisation (préférence pour les options extrêmes). Le format à choix forcé neutralise ces tendances en obligeant les répondants à faire un choix plus réfléchi entre deux alternatives.

Ces raisons soutiennent la décision d'opter pour un questionnaire à choix forcé, maximisant ainsi la validité et l'efficacité de DRIVE dans les processus d'évaluation professionnelle. L'utilisation du terme "choix forcé" plutôt que "format ipsatif" pour décrire le questionnaire DRIVE est liée aux spécificités psychométriques et à la finalité des scores obtenus. Le format ipsatif, caractérisé par une somme constante des scores, pose en effet plusieurs problèmes psychométriques :

- **Relativité des scores** : Dans un format ipsatif, les scores obtenus reflètent une norme personnelle, c'est-à-dire qu'ils indiquent la priorité qu'une personne donne à certaines caractéristiques par rapport à d'autres. Ainsi, ils ne représentent pas une mesure absolue, mais plutôt une comparaison relative entre différentes préférences ou tendances (Closs, 1996; Hicks, 1970; C. E. Johnson et al., 1988).
- **Validité de construit** : Les interdépendances entre les items dans un format ipsatif peuvent limiter la validité de construit. La covariance entre les items peut entraîner des distorsions statistiques, rendant difficile la distinction entre les construits (Baron, 1996; Cornwell & Dunlap, 1994).
- **Fidélité interne** : La fidélité, ou consistance interne, d'un test ipsatif est souvent questionnée. Les scores ipsatifs ne sont pas indépendants les uns des autres, ce qui complexifie la comparaison de la fidélité interne d'un format ipsatif à celle d'un format normatif (Meade, 2004).

En revanche, le format "choix forcé" employé par DRIVE vise à minimiser ces problèmes en n'imposant pas une somme constante des scores, permettant ainsi une évaluation plus nuancée des préférences individuelles sans les contraintes psychométriques associées au format ipsatif. Cela offre des avantages en termes de précision et de validité des mesures, ce qui est crucial pour les applications de sélection professionnelle où l'interprétation des scores doit être aussi fidèle que possible à la réalité des traits évalués. Aussi, en choisissant d'appliquer la Théorie de Réponse à l'Item (IRT) pour le calcul des résultats de DRIVE plutôt que la Théorie Classique des Tests (CTT), nous avons pris une décision

stratégique pour surmonter les limites intrinsèques des formats de questionnaire à choix forcé. L'IRT nous offre une mesure individualisée et précise, en considérant la difficulté spécifique de chaque item et la compétence du répondant. Cela permet d'obtenir des estimations des traits mesurés qui ne sont pas affectées par les variations de l'échantillon ou des items du test. De plus, la recherche menée par Brown & Maydeu-Olivares en 2012 appuie l'idée que l'IRT peut effectivement transcender les biais traditionnellement associés au format à choix forcé, offrant ainsi une capacité de comparaison plus équitable et des interprétations plus fiables des données. Cette approche avancée garantit que DRIVE fournit des évaluations des motivations non seulement sophistiquées, adaptées et solides.

5. Validité

Comment savoir si un questionnaire mesure réellement ce qu'il est censé mesuré ? Comment s'assurer de la bonne mesure de chaque échelle ainsi que la juste signification des résultats au questionnaire ? Les réponses à ces questions sont obtenues grâce à sa validation. L'objectif de la validation d'un questionnaire consiste à confirmer que celui-ci mesure réellement ce qu'il cherche à mesurer, et à quel point les résultats que nous pouvons en tirer son précis. Historiquement, la validité a été définie comme une corrélation entre un score à un questionnaire et un critère externe, qui mesure soit le même construit, ou qui mesure un construit censé être en lien avec le construit associé à ce score. Plusieurs types de validité sont nécessaires pour établir et assurer la validité d'un questionnaire. Les études de validité de DRIVE portent sur les types de validité suivants :

- **Validité de contenu** : comme son nom l'indique, la validité de contenu repose sur la nature sémantique du contenu de l'item par rapport au construit mesuré. En ce sens, le contenu doit être en rapport direct avec le construit qu'il est censé mesuré, et aussi couvrir tous les aspects principaux du construit mesuré ;
- **Validité de construit** : elle se réfère à la mesure dans laquelle le questionnaire évalue réellement la facette ou le construit psychologique qu'il est censé évaluer. Sont notamment proposées des analyses relatives à la saturation item-dimension, la corrélation inter-dimension, et aux paramètres de distribution ;
- **Validité prédictive** : la validité prédictive d'un questionnaire de motivations est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire de motivations peuvent être utilisés pour prédire la performance future.

5.1. Validité de contenu

La validité de contenu est un critère crucial dans l'élaboration de questionnaires psychométriques, qui mesure la pertinence du contenu et sa capacité à couvrir toutes les facettes d'un concept théorique. Elle évalue non seulement la représentativité des items par rapport au modèle théorique sous-jacent, mais aussi la qualité de leur formulation. Contrairement aux méthodes qui s'appuient sur des analyses statistiques, la validité de contenu exige une démarche rationnelle pour établir un lien entre le construit à mesurer et les questions posées. Une méthode commune pour apprécier cette validité consiste à recueillir l'avis d'experts dans le domaine ciblé par l'inventaire. Ces experts jugent la pertinence des items et peuvent également être invités à associer chaque item aux dimensions théoriques qu'ils représentent, sans préconceptions. Cette approche, qui a été privilégiée par l'équipe Science d'AssessFirst, offre un processus de validation plus objectif et informatif. Cette méthode a été spécifiquement choisie lors du développement de DRIVE car elle correspondait le mieux aux standards et aux exigences de l'époque, assurant ainsi une base solide et fiable pour l'outil.

L'étude de validité de contenu de DRIVE a été réalisée avec dix juges qui ont été sélectionnés pour participer à cette étude. Chaque juge a reçu un premier document (document A) détaillant les 180 items de l'inventaire DRIVE ainsi qu'un second document (document B) présentant les définitions des 20 dimensions évaluées par cet inventaire. La consigne présentée aux juges était la suivante :

> *L'inventaire DRIVE comporte 20 dimensions. Chacune des dimensions comporte 9 items. Pour chacun des items présentés dans le document A, vous devez sélectionner dans le document B la dimension à laquelle il se rapporte. Un item ne peut appartenir qu'à une unique facette.*

Le tableau suivant présente, pour chaque dimension, le nombre d'items (sur les 9 items de cette dimension) dont l'accord inter-juges est très élevé (entre 90 et 100 %), élevé (entre 80 et 90 %), correct (entre 70 et 80 %) et insuffisant (entre 0 et 70 %). L'étude menée sur la validité de contenu de DRIVE révèle des résultats très positifs. L'accord inter-juges – qui mesure le consensus entre experts sur la pertinence des items par rapport aux dimensions qu'ils sont censés mesurer – est extrêmement encourageant. Avec un taux très élevé d'accord inter-juges pour la majorité des dimensions (7,25 sur les dimensions évaluées), et un niveau élevé pour une petite proportion (1,05 dimension), cela indique une forte cohérence dans l'évaluation des experts. De plus, le niveau de concordance est considéré comme correct pour moins d'une dimension (0,7 dimension), sans qu'aucun item n'ait été jugé comme ayant un degré d'accord insuffisant. Ces données démontrent que les items de DRIVE sont perçus de manière quasi-unanime comme étant pertinents et représentatifs des dimensions qu'ils sont destinés à mesurer. Ainsi, on peut affirmer avec assurance que la validité de contenu de DRIVE est robuste, reflétant une construction rigoureuse et une méthodologie fiable qui renforcent sa crédibilité et sa valeur en tant qu'outil d'évaluation psychométrique.

Dimensions	91-100%	81-90%	71-80%	0-70%	Total
Créer de nouvelles choses	9	0	0	0	9
Se dépasser au quotidien	8	1	0	0	9
Se préoccuper de l'esthétique	7	1	1	0	9
Analyser des données	7	1	1	0	9
Rencontrer de nouvelles personnes	7	1	1	0	9
Avoir des tâches clairement définies	7	2	0	0	9
Se préoccuper de la qualité	6	2	1	0	9
Avoir de l'influence	8	0	1	0	9
Avoir de l'autonomie	9	0	0	0	9
Travailler en équipe	8	1	0	0	9
Avoir un impact positif sur le monde	7	2	0	0	9
Travailler dans une ambiance fun	7	0	2	0	9
Evoluer dans un environnement sécurisant	7	1	1	0	9
Travailler de façon disciplinée	8	0	1	0	9
Préserver son équilibre personnel	8	1	0	0	9
Recevoir des récompenses	6	2	1	0	9
Avoir une rémunération attractive	5	2	2	0	9
Remporter régulièrement des succès	8	1	0	0	9

Table 5.1. Taux d'accord inter-juges pour chaque dimension.

Aider les autres	6	2	1	0	9
Être reconnu(e) par les autres	7	1	1	0	9
Moyenne du degré d'accord	**7.25**	**1.05**	**.70**	**0**	**9**

5.2. Validité de construit

La validité de construit permet de savoir si le test mesure réellement le construit théorique souhaité, ou autre chose. Ce type de validité chevauche certains des autres aspects de la validité, car toute preuve de validité contribue à la compréhension de la validité de construit d'un instrument d'évaluation. La validité de construit est importante, car elle influe sur l'interprétation des scores d'un questionnaire. Si un questionnaire prétend mesurer une dimension de motivation spécifique, comment être sûr qu'il mesure réellement cette dimension ? Si un questionnaire est censé mesurer une dimension, mais qu'en réalité, ce n'est pas le cas, toute interprétation du score est incorrecte et pourrait conduire à des décisions biaisées. La validité de construit ne cherche pas simplement à savoir si un questionnaire mesure une dimension. Aussi, elle considère des investigations cherchant à savoir si les interprétations des résultats des tests sont cohérentes avec un réseau impliquant des termes théoriques et d'observation (Cronbach & Meehl, 1955).

Il n'existe pas de méthode unique pour déterminer la validité de construit, mais différentes méthodes et approches sont combinées. Pour évaluer la validité de construit de DRIVE, nous privilégions ici trois méthodes complémentaires, à savoir la saturation item-dimension et la corrélation inter-dimensions (Thurstone, 1947; Bollen, 1989; McDonald, 2013), ainsi que les paramètres de distribution.

- la **saturation item-dimension** fait référence à la corrélation entre les scores d'un item et les scores totaux de la dimension ou du facteur auquel il est censé appartenir. En d'autres termes, si un item est censé mesurer une dimension spécifique, alors il devrait être étroitement associé aux autres items qui mesurent cette dimension. Ainsi, plus la corrélation entre un item et la dimension est élevée, plus l'item est considéré comme étant fortement lié à cette dimension et donc plus valide. Pour la saturation item-dimension, une valeur supérieure ou égale .40 est généralement considérée comme satisfaisante et adéquate : cela suggère que l'item mesure bien la dimension à laquelle il est censé se rapporter (Campbell & Fiske, 1959; Nunnally, 1978; Hair, Black, Babin & Anderson, 1998). Une saturation inférieure à .40 peut être acceptable si elle est soutenue par une justification théorique.

- la **corrélation inter-dimensions** évalue la relation entre les scores de différents facteurs ou dimensions mesurées par un test. Si deux dimensions sont censées être distinctes et indépendantes, alors elles devraient avoir des scores faiblement corrélés. En revanche, si les dimensions sont étroitement liées ou si elles se chevauchent, les scores devraient être plus fortement corrélés. Il n'existe pas de seuil universel pour la corrélation inter-dimension. Toutefois, certains auteurs suggèrent que la corrélation inter-dimensions ne devrait pas dépasser la racine carrée de la variance de chaque dimension pour établir la validité discriminante (Hair, Black, Babin & Anderson, 1998). Il est généralement souhaitable que les dimensions soient relativement indépendantes, bien qu'il puisse exister certaines corrélations modérées entre les dimensions qui sont justifiées par le modèle théorique sous-jacent.

- **les paramètres de distribution** correspondent à la distribution des scores dans les questionnaires. Ils permettent d'identifier les scores atypiques, d'explorer les différences individuelles dans la distribution des scores, et de mieux interpréter les résultats.

5.2.1. Saturation item-dimension

Les dernières études de saturation item-dimension de DRIVE ont été conduites en 2017 (N = 4450) sur les items principaux mesurant chaque dimension. Le tableau ci-dessous présente les résultats de cette analyse : (1) la saturation varie de .52 ≤ r ≤ .83, (2) la saturation moyenne par facette varie de .63 ≤ r ≤ .69. Ces conclusions permettent ainsi d'attester de saturations item-dimension satisfaisantes et adéquates.

Dimensions	i1	i2	i3	i4	i5	i6	i7	i8	i9	Moyenne
Créer de nouvelles choses	.69	.59	.65	.71	.71	.65	.78	.62	.60	**.67**
Se dépasser au quotidien	.63	.64	.73	.60	.76	.66	.56	.61	.68	**.65**
Se préoccuper de l'esthétique	.60	.63	.58	.66	.65	.60	.647	.72	.62	**.63**
Analyser des données	.57	.83	.76	.67	.71	.65	.63	.63	.65	**.68**
Rencontrer de nouvelles personnes	.70	.69	.7	.62	.65	.60	.64	.64	.78	**.67**
Avoir des tâches clairement définies	.68	.67	.61	.76	.69	.75	.71	.61	.61	**.68**
Se préoccuper de la qualité	.65	.64	.64	.73	.61	.59	.60	.71	.59	**.64**
Avoir de l'influence	.71	.65	.64	.72	.61	.66	.73	.62	.57	**.66**
Avoir de l'autonomie	.61	.69	.61	.67	.66	.60	.68	.70	.67	**.65**
Travailler en équipe	.67	.7	.75	.68	.71	.68	.69	.65	.66	**.69**
Avoir un impact positif sur le monde	.63	.66	.61	.65	.67	.52	.77	.63	.62	**.64**
Travailler dans une ambiance fun	.65	.64	.71	.61	.66	.67	.74	.68	.67	**.67**
Evoluer dans un environnement sécurisant	.62	.57	.74	.62	.62	.61	.74	.60	.70	**.65**
Préserver son équilibre personnel	.60	.65	.60	.63	.67	.59	.67	.61	.68	**.63**
Travailler de façon disciplinée	.65	.78	.62	.60	.67	.63	.61	.67	.75	**.66**
Recevoir des récompenses	.66	.67	.67	.75	.64	.65	.59	.69	.65	**.66**
Avoir une rémunération attractive	.69	.7	.61	.7	.68	.62	.64	.61	.71	**.66**

Table 5.2. Saturation item-dimension pour chaque item.

5.2.2. Corrélation inter-dimension

Les dernières études de corrélation inter-dimension de DRIVE ont été conduites en 2017 (N = 4450). Les dimensions sont faiblement corrélées, ce qui permet de justifier d'un niveau de consistance acceptable. Les dimensions les plus corrélées sont les paires « Avoir une rémunération attractive » et « Avoir un impact positif sur le monde » (r = -.45), « Avoir une rémunération attractive » et « Recevoir des récompenses » (r = .43), et « Avoir des tâches clairement définies » et « Avoir de l'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. Corrélations inter-dimensions des échelles DRIVE.

5.2.3. Paramètres de distribution

La distribution des scores obtenus par un questionnaire est un aspect important de la validité de construit du questionnaire. En effet, la manière dont les scores sont distribués pour chaque dimension de motivation peut fournir des informations importantes sur la manière dont le test mesure réellement cette dimension, ainsi que sur la manière dont les scores sont interprétés. Nos analyses de paramètres de distribution se centrent sur 5 paramètres principaux et nécessaires :

- la **moyenne**, qui est un indicateur de la tendance centrale des scores dans la distribution ;
- la **médiane**, qui est la valeur qui divise la distribution en deux parties égales : la moitié des scores sont supérieurs à la médiane, et l'autre moitié sont inférieurs. C'est également un indicateur de la tendance centrale des scores dans la distribution, qui est moins sensible aux scores extrêmes que la moyenne ;
- l'**écart-type**, qui est une mesure de la dispersion des scores autour de la moyenne. Il est calculé en prenant la racine carrée de la variance des scores ;
- l'**asymétrie** , qui est calculée en comparant la fréquence des scores à gauche et à droite de la moyenne. Si la distribution est parfaitement symétrique, l'asymétrie est nulle. Si la distribution est asymétrique vers la gauche, l'asymétrie est négative. Si la distribution est asymétrique vers la droite, l'asymétrie est positive ;
- l'**aplatissement** (ou coefficient d'aplatissement), qui est calculé en comparant la distribution des scores à une distribution normale. Si la distribution est moins aplatie que la distribution normale, l'aplatissement est négatif. Si la distribution est plus aplatie que la distribution normale, l'aplatissement est positif.

Les attentes pour les paramètres de distribution dépendent du contexte et de l'instrument de mesure utilisé. Toutefois, en général, voici ce que l'on attend pour de « bons » paramètres de distribution :

- La moyenne doit être proche de la valeur médiane : cela indique que la distribution est symétrique. Si la moyenne est significativement différente de la médiane, cela peut indiquer une asymétrie de distribution ;

- L'écart-type doit être raisonnable et suffisamment grand pour capturer les différences individuelles dans la dimension mesurée, mais pas trop grand pour ne pas diluer les différences entre les individus. En général, on s'attend à ce que l'écart-type soit d'environ 2 pour les échelles de personnalité en 10 points ;
- L'asymétrie (skewness) doit être proche de 0 (distribution symétrique). Si l'asymétrie est significativement différente de 0, cela peut indiquer une asymétrie dans la distribution ;
- L'aplatissement (kurtosis) doit être proche de 0 (distribution normale). Si l'aplatissement est significativement différent de 0, c'est que la distribution est soit plus plate, soit plus pointue.

Les dernières études de paramètres de distribution de DRIVE ont été conduites en 2023 (N = 48449).

Dimension	Moyenne	Médiane	Écart-type	Asymétrie	Aplatissement
Créer de nouvelles choses	5.71	6.00	2.02	0.05	-0.25
Se dépasser au quotidien	5.45	5.00	1.93	0.17	0.21
Se préoccuper de l'esthétique	5.33	5.00	1.98	-0.10	0.05
Analyser des données	5.45	5.00	2.03	0.04	-0.66
Rencontrer de nouvelles personnes	5.65	6.00	2.03	-0.04	-0.59
Avoir des tâches clairement définies	5.27	5.00	1.84	0.02	-0.38
Se préoccuper de la qualité	5.10	5.00	1.91	0.12	-0.15
Avoir de l'influence	5.69	6.00	1.90	-0.04	-0.32
Avoir de l'autonomie	5.52	6.00	1.86	-0.15	-0.08
Travailler en équipe	5.71	6.00	1.92	0.09	0.14
Avoir un impact positif sur le monde	5.93	6.00	2.09	0.00	-0.43
Travailler dans une ambiance fun	5.92	6.00	1.90	-0.05	-0.33
Evoluer dans un environnement sécurisant	5.12	5.00	2.01	-0.01	-0.43
Travailler de façon disciplinée	5.52	6.00	1.96	-0.10	-0.24
Préserver son équilibre personnel	5.12	5.00	1.89	-0.05	-0.22
Recevoir des récompenses	5.29	5.00	2.04	-0.15	-0.27
Avoir une rémunération attractive	5.17	5.00	2.06	0.10	-0.70
Remporter régulièrement des succès	5.33	5.00	1.82	0.10	-0.13
Aider les autres	5.60	6.00	1.90	0.12	0.26
Être reconnu(e) par les autres	5.69	6.00	1.83	0.07	-0.24
	5.48	*5.50*	*1.95*	*.01*	*-.24*

Table 5.4. Paramètres de distribution des échelles DRIVE.

Les paramètres de distribution sont ainsi cohérents avec les standards attendus. La normalité des distributions peut être interprétée par des indices comme la superposition des moyennes et des médianes, et à l'aide des coefficients de symétrie et d'aplatissement qui sont proches de 0.

5.3. Validité prédictive

La validité prédictive d'un questionnaire de motivations est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire peuvent être utilisés pour prédire la performance future. La preuve de validité prédictive est

particulièrement applicable lorsque l'on souhaite faire une inférence, à partir d'un score du questionnaire, de la position de l'évalué sur une autre variable de critère évaluée de manière indépendante à une date ultérieure. Les études de validité prédictive liées à DRIVE sont présentées dans un manuel dédié.

5.4. Conclusion

La validation d'un questionnaire de motivations est cruciale pour s'assurer que les mesures obtenues sont précises. Dans cette étude, nous avons examiné la validité de contenu, la validité de construit, et la validité prédictive de DRIVE. Nos analyses montrent que : (1) DRIVE couvre les construits théoriques qu'il est censé mesurer, (2) le questionnaire est bien structuré et montre une bonne homogénéité de la mesure, (3) les dimensions sont prédictives de la réussite en poste. Dans l'ensemble, les résultats obtenus répondent aux standards psychométriques les plus exigeants, et démontrent la validité de DRIVE. Aller plus loin dans les qualités psychométriques de DRIVE requiert toutefois de s'intéresser à sa fidélité. En ce sens, un questionnaire doit être valide et fidèle pour être utilisé dans le cadre de décisions professionnelles (recrutement, mobilité, etc.). En effet, un questionnaire valide mais non fidèle signifierait que le test mesure bien ce qu'il doit mesurer, mais que les scores individuels sont incohérents. Au contraire, un questionnaire valide et fidèle signifie que le questionnaire mesure systématiquement ce qu'il est censé mesurer : en d'autres mots, il frappe constamment dans le mille. Les preuves de fidélité sont présentées dans le chapitre suivant.

6. Fidélité

Comment savoir si les résultats d'un questionnaire sont fiables ? Comment s'assurer que le questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ? Les réponses à ces questions sont obtenues grâce à l'étude de sa fidélité. Alors que la validité renseigne sur la capacité d'un questionnaire à réellement mesurer ce que l'on souhaite mesurer, la fidélité permet de savoir si cette mesure sera consistante et fiable à chaque fois que ce même questionnaire est complété par une même personne. En somme, la fidélité d'un questionnaire est une mesure de sa cohérence ou de sa stabilité dans le temps. Elle vise à déterminer si un questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ou à des personnes similaires. L'objectif de la fidélité est donc de garantir que les résultats obtenus sont fiables et précis. La fidélité d'un questionnaire peut s'évaluer de deux manières différentes et complémentaires :

- **La cohérence interne,** qui est une mesure statistique utilisée pour évaluer la fidélité d'un test psychométrique. Elle évalue l'homogénéité ou la similarité des différents items d'un test qui sont supposés mesurer la même dimension psychologique. Autrement dit, la cohérence interne évalue si plusieurs éléments qui proposent de mesurer la même chose produisent des scores similaires ;

- **La fidélité test-retest,** qui est une méthode pour évaluer la fidélité d'une mesure en mesurant la même variable à deux moments différents. Elle permet de mesurer la stabilité temporelle de la mesure et d'estimer la proportion de la variance totale qui est attribuable à l'erreur de mesure. Elle est souvent utilisée dans les études longitudinales ou pour évaluer la stabilité d'un test sur une période de temps donnée.

6.1. Cohérence interne

Les dernières études d'alpha de Cronbach pour DRIVE ont été conduites en 2015 (N = 332). À l'exception de 4 dimensions, qui présentent des α compris entre .62 et .69, les coefficients α sont tous supérieurs à .70, montrant des résultats de fidélité adéquats du questionnaire DRIVE. Qui plus est, ces résultats sont à mettre en perspective au regard de la structure à choix forcé de DRIVE, qui fait de l'α de Cronbach un indicateur dont le pertinence n'est ici pas optimale.

Dimension	α	Dimension	α
Créer de nouvelles choses	.77	Avoir un impact positif sur le monde	.72
Se dépasser au quotidien	.71	Travailler dans une ambiance fun	.72
Se préoccuper de l'esthétique	.68	Evoluer dans un environnement sécurisant	.72
Analyser des données	.70	Travailler de façon disciplinée	.77
Rencontrer de nouvelles personnes	.75	Préserver son équilibre personnel	.81
Avoir des tâches clairement définies	.87	Recevoir des récompenses	.71
Se préoccuper de la qualité	.77	Avoir une rémunération attractive	.75
Avoir de l'influence	.74	Remporter régulièrement des succès	.82
Avoir de l'autonomie	.69	Aider les autres	.62
Travailler en équipe	.79	Être reconnu(e) par les autres	.67

Table 6.1. Coefficients alpha des échelles DRIVE.

6.2. Fidélité test-retest

La fidélité test-retest est une mesure de la cohérence temporelle d'un questionnaire ou d'une échelle de mesure. Elle consiste à administrer le même questionnaire à un groupe de participants à deux moments différents, avec un intervalle de temps entre les deux administrations. La corrélation entre les deux résultats est ensuite calculée pour déterminer la fidélité du test. Si la corrélation est élevée, cela signifie que les scores des participants sont stables et que le questionnaire peut donc être considéré comme fiable.

Les dernières études de fidélité test-retest pour DRIVE ont été conduites en 2017 (N = 232). Le tableau ci-dessous fournit la fidélité test-retest de DRIVE administré à 3 mois d'intervalle. Les 20 dimensions du questionnaire DRIVE démontrent une bonne fidélité test-retest avec des coefficients allant de .67 à .83 et un coefficient de fidélité médian de .74. Cette stabilité des scores sur une période significative souligne la capacité de DRIVE à fournir des mesures cohérentes et fiables des motivations, un atout essentiel pour les évaluations en contexte professionnel où les décisions à long terme se basent sur ces données.

Dimension	Pearson r	Dimension	Pearson r
Créer de nouvelles choses	.76	Avoir un impact positif sur le monde	.71
Se dépasser au quotidien	.72	Travailler dans une ambiance fun	.77
Se préoccuper de l'esthétique	.69	Evoluer dans un environnement sécurisant	.70
Analyser des données	.73	Travailler de façon disciplinée	.69
Rencontrer de nouvelles personnes	.75	Préserver son équilibre personnel	.83
Avoir des tâches clairement définies	.79	Recevoir des récompenses	.74
Se préoccuper de la qualité	.74	Avoir une rémunération attractive	.71
Avoir de l'influence	.77	Remporter régulièrement des succès	.75
Avoir de l'autonomie	.67	Aider les autres	.77
Travailler en équipe	.74	Être reconnu(e) par les autres	.71

Table 6.2. Fidélité test-retest des échelles DRIVE.

7. Sensibilité

La sensibilité, aussi appelée discrimination, désigne la capacité d'un questionnaire à distinguer les personnes ayant un niveau élevé sur une facette et les personnes ayant un niveau faible. Elle reflète donc la capacité du questionnaire à identifier la singularité de chaque individu. Un questionnaire de motivations sensible peut ainsi identifier les différences subtiles entre les personnes, et aider à comprendre leurs comportements avec davantage de précision et de discrimination. La sensibilité se réfère ainsi à la capacité du questionnaire à identifier correctement les personnes qui possèdent la caractéristique mesurée et à éviter les faux positifs (personnes qui ne possèdent pas la caractéristique, mais qui sont identifiées comme telles par le questionnaire). Cette propriété est directement liée à la qualité du score.

Les premières tentatives de mesure de la discrimination étaient basées sur des échelles cumulatives (Guttman, 1944; Walker, 1931; Loevinger, 1948; Loevinger, 1953, cités par Hankins, 2008). Ferguson est toutefois l'un des premiers à proposer de conceptualiser la discrimination sous la forme d'un coefficient. En ce sens, s'il existe un nombre maximum de différences possibles dans un échantillon, le coefficient de discrimination correspond au ratio entre le nombre de différences réellement constatées, et ce nombre maximal de différences. Ce coefficient, appelé delta de Ferguson (Ferguson, 1949; Kline, 2000), est ainsi le rapport entre les différences observées entre les personnes et le nombre de différences maximum possibles. Il se veut être un indice direct et non-paramétrique du degré de distinction faite par un instrument entre des individus. Si aucune différence n'est observée, alors $\delta = 0$. Si toutes les discriminations possibles sont observées, alors $\delta = 1$. Généralement, une distribution normale devait avoir une excellente discrimination, où $\delta \geqslant .9$ (Ferguson, 1949). Les discriminations plus faibles étant attendues pour les distributions leptokurtiques (car ces distributions ne parviennent pas à discriminer autour de la moyenne) et les distributions asymétriques (car celles-ci ne parviennent pas à discriminer à une extrémité de la distribution). Démontrer l'excellente discrimination des échelles d'un questionnaire requiert donc un $\delta \geqslant .9$.

Les dernières études sensibilité de DRIVE, basées sur le calcul du δ de Ferguson, ont été conduites en 2023 (N = 12386). Les coefficients δ sont tous supérieurs à .9, montrant une excellente discrimination des échelles de mesure. En d'autres termes, cela signifie que le questionnaire est capable de détecter avec précision les différences individuelles dans les motivations des personnes testées, et qu'il est sensible aux variations individuelles dans les dimensions mesurées. Les résultats sont présentés dans le tableau ci-dessous.

Dimensions	δ	Dimensions	δ
Créer de nouvelles choses	.97	Avoir un impact positif sur le monde	.98
Se dépasser au quotidien	.97	Travailler dans une ambiance fun	.96
Se préoccuper de l'esthétique	.97	Evoluer dans un environnement sécurisant	.96
Analyser des données	.97	Travailler de façon disciplinée	.98
Rencontrer de nouvelles personnes	.97	Préserver son équilibre personnel	.97
Avoir des tâches clairement définies	.96	Recevoir des récompenses	.99
Se préoccuper de la qualité	.97	Avoir une rémunération attractive	.98
Avoir de l'influence	.96	Remporter régulièrement des succès	.94
Avoir de l'autonomie	.97	Aider les autres	.98
Travailler en équipe	.98	Être reconnu(e) par les autres	.96

Table 7.1. Delta de Ferguson des échelles DRIVE.

8. Équité

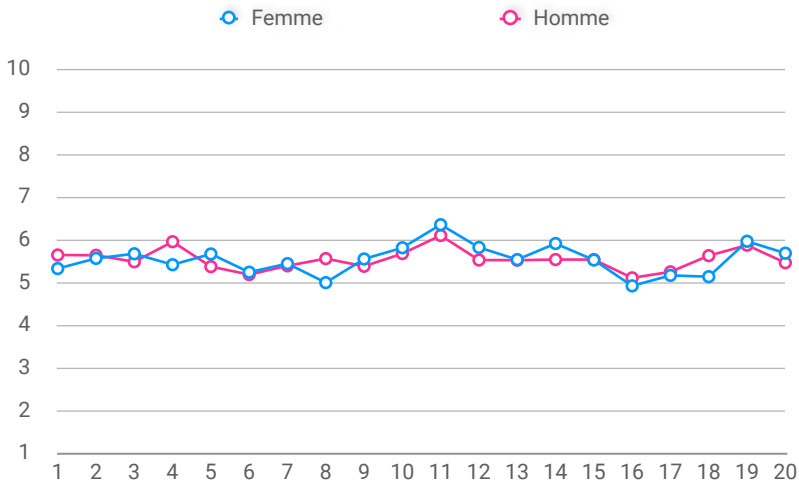
L'équité dans le contexte d'un questionnaire de motivations fait référence à la mesure dans laquelle celui-ci est équitable et impartial pour toutes les personnes qui le passent, quelle que soit leur origine, leur sexe, leur orientation sexuelle, leur race ou leur culture. En d'autres termes, un questionnaire équitable est conçu pour être objectif et juste pour toutes les personnes qui le passent, sans biais ou discrimination à l'encontre d'un groupe particulier. Nos équipes mettent en ce sens tout en œuvre pour veiller au caractère équitable des questionnaires et des analyses prédictives, et pour s'assurer que l'usage de nos algorithmes dans les processus décisionnels n'amènent pas de discriminations via des biais algorithmiques insoupçonnés.

8.1. Équité dans les résultats

Les données présentées dans cette section démontrent qu'il n'existe pas de différences majeures ou de taille d'effet forte dans les résultats au questionnaire DRIVE en fonction des variables de genre et d'âge. Note : seules les informations personnelles nécessaires à l'utilisation correcte d'AssessFirst sont demandées à nos utilisateurs. Par exemple, il n'y a aucune mention de l'orientation religieuse, politique ou sexuelle, quel que soit le moment. Quand il s'agit de l'âge, nous demandons la date de naissance afin de nous assurer que cela n'affecte pas la façon dont les questions sont traitées. De même, toutes les variables ci-après analysées n'interviennent à aucun moment dans le calcul des résultats dans la solution.

8.1.1. Équité selon le genre

Les dernières analyses d'équité selon le genre pour DRIVE ont été conduites en 2022, sur un échantillon (N = 332587) composé de 51% d'hommes et 49% de femmes.



Graphique 8.1. Scores moyens aux 20 dimensions DRIVE en fonction du genre.

Globalement, toutes les moyennes se situent entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Il n'existe donc pas de différence majeure entre les résultats des hommes et ceux des femmes sur les 20 dimensions mesurées par DRIVE. Les résultats sont ainsi équitables selon le genre. Ces conclusions sont par ailleurs étayées par la taille des effets, présentée ci-après.

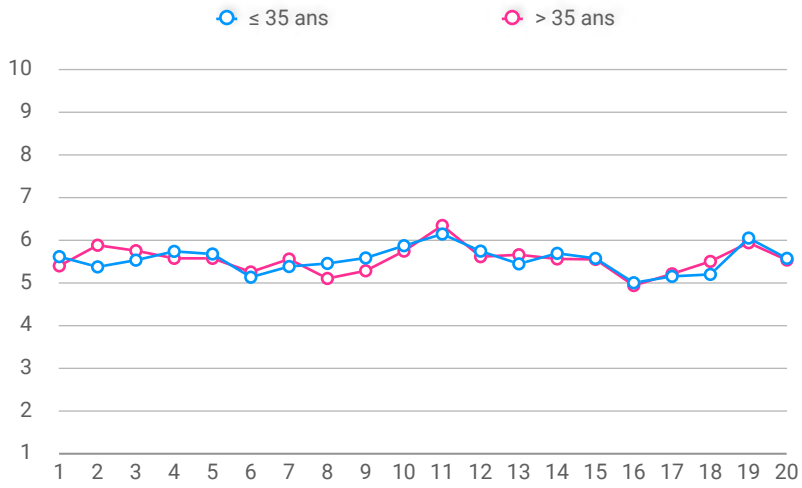
Les tailles d'effet ici mises en lumière viennent supporter la conclusion précédente. Il convient toutefois de noter que : (1) ces effets sont rares, (2) il s'agit exclusivement d'effets très faibles, (3) ils sont potentiellement inhérents à un effet d'échantillonnage. En conclusion, bien que quelques effets soient mis en avant, il n'existe toutefois pas de différence majeure entre les résultats des hommes et des femmes sur les 20 dimensions DRIVE. Les résultats de DRIVE sont ainsi équitables selon le genre.

Dimensions	*d*	Effet	Dimensions	*d*	Effet
Créer de nouvelles choses	.16	-	Avoir un impact positif sur le monde	-.12	-
Se dépasser au quotidien	.04	-	Travailler dans une ambiance fun	-.15	-
Se préoccuper de l'esthétique	-.09	-	Evoluer dans un environnement sécurisant	-.01	-
Analyser des données	.26	Très faible	Travailler de façon disciplinée	-.18	-
Rencontrer de nouvelles personnes	-.15	-	Préserver son équilibre personnel	.00	-
Avoir des tâches clairement définies	-.03	-	Recevoir des récompenses	.08	-
Se préoccuper de la qualité	-.03	-	Avoir une rémunération attractive	.04	-
Avoir de l'influence	.29	Très faible	Remporter régulièrement des succès	.27	Très faible
Avoir de l'autonomie	-.09	-	Aider les autres	-.04	-
Travailler en équipe	-.06	-	Être reconnu(e) par les autres	-.12	-

Table 8.1. d de Cohen pour le genre des échelles DRIVE.

8.1.2. Équité selon l'âge

Les dernières analyses d'équité selon l'âge pour DRIVE ont été conduites en 2022, sur un échantillon (N = 64675) composé de 42% d'individus de moins de 35 ans, et 58% d'individus de plus de 35 ans. Pour comparer les résultats, nous avons découpé l'échantillon en deux classes d'âge à partir de l'âge de 35 ans (en raison de l'âge moyen de l'échantillon égal à 33.3 ans - avec un écart type de 10.2 ans).



Graphique 8.2. Scores moyens aux 20 dimensions DRIVE en fonction de l'âge.

Globalement, toutes les moyennes se situent entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Il n'existe donc pas de différence majeure entre les résultats des personnes en fonction de leur âge sur 20 dimensions mesurées par DRIVE. Les résultats sont ainsi équitables selon l'âge. Ces conclusions sont par ailleurs étayées par la taille des effets, présentée ci-après.

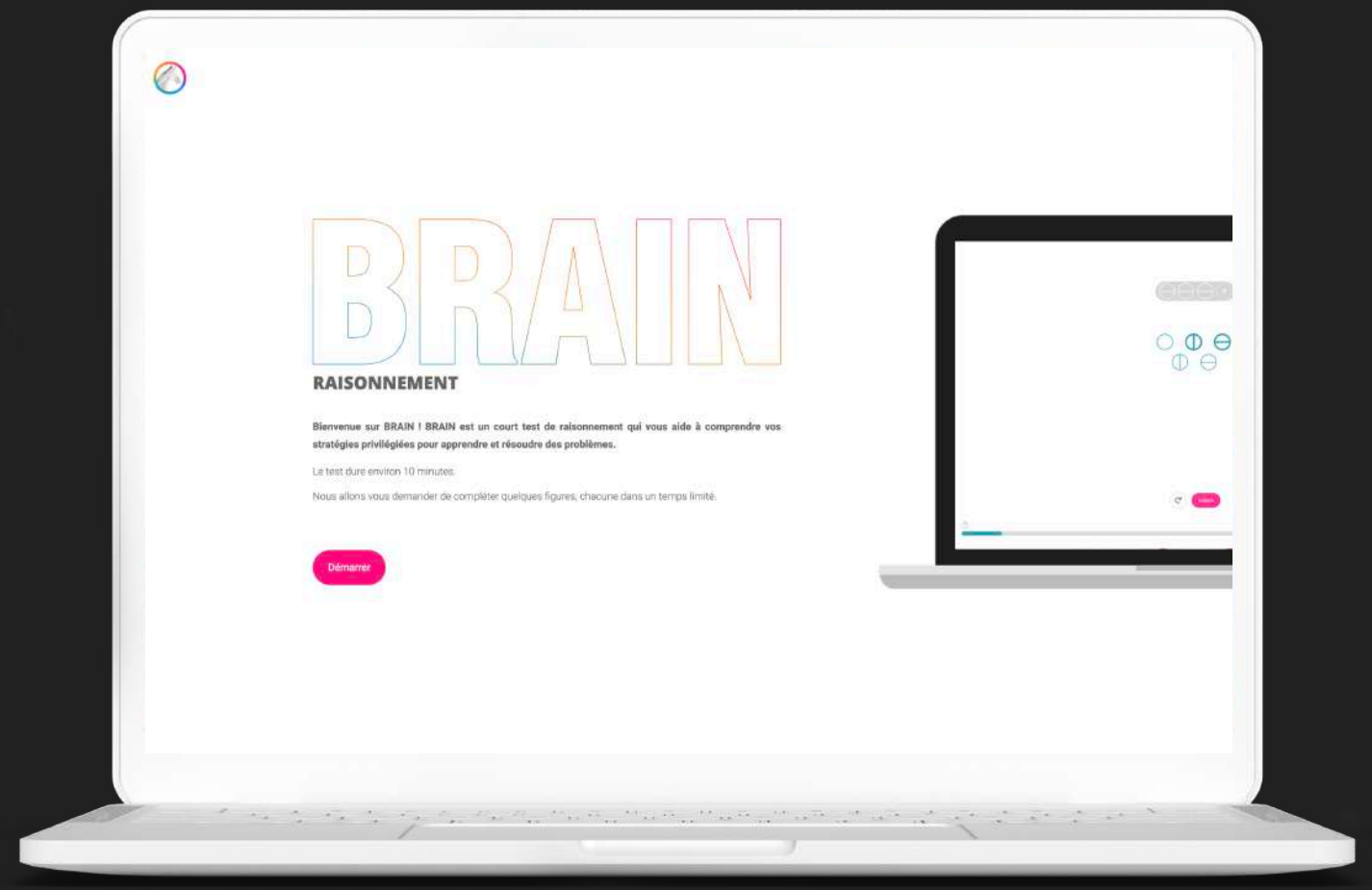
Les tailles d'effet ici mises en lumière viennent supporter la conclusion précédente. Il convient toutefois de noter que : (1) ces effets sont rares, (2) il s'agit exclusivement d'effets très faibles, (3) ils sont potentiellement inhérents à un effet d'échantillonnage. En conclusion, bien que quelques effets soient mis en avant, il n'existe toutefois pas de différence majeure entre les 2 catégories d'âges sur les 20 dimensions DRIVE. Les résultats de DRIVE sont ainsi équitables selon l'âge.

Dimensions	d	.09	Dimensions	d	Effet
Créer de nouvelles choses	.11	-	Avoir un impact positif sur le monde	.09	-
Se dépasser au quotidien	.25	Très faible	Travailler dans une ambiance fun	.07	-
Se préoccuper de l'esthétique	.11	-	Evoluer dans un environnement sécurisant	.11	-
Analyser des données	.08	-	Travailler de façon disciplinée	.06	-
Rencontrer de nouvelles personnes	.05	-	Préserver son équilibre personnel	.01	-
Avoir des tâches clairement définies	.06	-	Recevoir des récompenses	.03	-
Se préoccuper de la qualité	.09	-	Avoir une rémunération attractive	.03	-
Avoir de l'influence	.18	-	Remporter régulièrement des succès	.16	-
Avoir de l'autonomie	.15	-	Aider les autres	.05	-
Travailler en équipe	.06	-	Être reconnu(e) par les autres	.02	-

Table 8.2. d de Cohen pour l'âge des échelles DRIVE.

8.1.3. Conclusion

Que cela soit en termes de genre ou de catégories d'âge, il n'existe pas de différences majeures entre les résultats obtenus par les différents groupes au questionnaire DRIVE d'AssessFirst. En somme, les résultats obtenus permettent de soutenir l'hypothèse selon laquelle le questionnaire proposé ne discrimine aucun public et s'avère équitable. Les effets les plus marqués concernent le genre et uniquement quelques facettes. Aussi, ces effets mentionnés sont très faibles.



Plongez dans la résolution de problèmes

BRAIN est un court questionnaire de raisonnement, qui permet de comprendre la capacité générale de raisonnement d'une personne et la manière dont elle prend des décisions. Adaptatif, gamifié et réalisable en 10 minutes, BRAIN permet de concilier fidélité de la mesure et qualité de l'expérience utilisateur.

BRAIN

1. Introduction

BRAIN est un court questionnaire de raisonnement, qui permet de comprendre la capacité générale de raisonnement d'une personne et la manière dont elle prend des décisions. Adaptatif, gamifié et réalisable en 10 minutes, BRAIN permet de concilier fidélité de la mesure et qualité de l'expérience utilisateur. Ce questionnaire est composé de **8 à 12 items** mesurant **plusieurs dimensions, parmi lesquelles :** score global de potentiel de raisonnement, vitesse de décision, type de tâches préférées, mode d'apprentissage, capacité à gérer la complexité, précision des décisions, facteurs de risques, style comportemental. BRAIN a été conçu en 2020 par l'équipe Science d'AssessFirst.

2. Historique de développement

Les avancées technologiques rendent le travail éphémère et complexe. La plupart des emplois exigent maintenant des compétences pour résoudre des problèmes non routiniers et dynamiques. À ce titre, les capacités de raisonnement d'une personne sont cruciales. L'intelligence se définit comme une capacité générale à comprendre et apprendre rapidement. Linda Gottfredson, professeure de psychologie à l'Université du Delaware, explique qu'il s'agit de la faculté à raisonner, à anticiper ou encore à penser des idées abstraites : en résumé, une facilité pour comprendre la complexité de l'environnement et y répondre. D'autres théories invitent néanmoins à considérer l'intelligence à travers des compétences spécifiques et multiples, en expliquant qu'un individu peut être performant dans un domaine (par exemple pour manipuler des chiffres), et moins bon sur d'autres exercices : Joy Paul Guilford, psychologue et professeur de psychologie à l'Université de Caroline du Sud, va jusqu'à distinguer 150 formes d'intelligences différentes. Dans cette controverse, les études ont depuis longtemps démontré que, même s'il existe des formes de raisonnement spécifiques, celles-ci sont fortement liées entre elles et à un facteur général d'intelligence, appelé facteur *g*. Au travail, le facteur *g* est l'une des variables les plus explicatives de la performance d'un collaborateur, qui plus est sur des fonctions impliquant des hauts niveaux de complexité : son importance semble en effet moindre dans des emplois où les prises de décisions et les problèmes sont plus simples et routiniers. A travers l'analyse de 20000 études regroupant 5 millions d'individus, Nathan Kuncel, Deniz Ones et Paul Sackett, chercheurs à l'Université du Minnesota, démontrent ainsi que les capacités cognitives permettent de prédire 25% de la performance future d'un collaborateur, notamment car elles leur permettent d'acquérir plus rapidement les connaissances nécessaires pour réaliser la mission (Kuncel, Ones and Sackett, 2010). Steffanie Wilk, Laura Desmarais et Paul Sackett, chercheurs en psychologie et management, concluent par ailleurs que les individus avec de fortes capacités cognitives évoluent plus facilement dans la hiérarchie (Wilk, Desmarais and Sackett, 1995).

Malgré ces résultats, les décisionnaires RH ont encore trop souvent tendance à valoriser d'autres critères, comme l'expérience, dans le recrutement des candidats ou dans la gestion des évolutions de carrière. Pourtant, John Hunter, professeur de psychologie à Michigan State, démontre que le niveau de raisonnement d'un individu est, en moyenne, trois fois plus prédictif de sa performance future que ne l'est son expérience. Une étude plus récente suggère même que l'expérience ne présente aucune corrélation significative avec la réussite en poste des nouveaux collaborateurs (Van Iddekinge, Frieder and Roth, 2019). Dans un monde du travail en disruption, la majorité des individus doit apprendre et s'approprier de nouvelles méthodes de production. Instaurer une politique de recrutement qui prend en compte les capacités de raisonnement des candidats est donc un pré-requis pour optimiser la performance des collaborateurs, leur adaptation au monde de demain et la croissance économique. Même si elle peut sembler élitiste, cette vision n'écarte toutefois pas de l'emploi les personnes avec des facultés cognitives moins importantes. En effet, bien que l'automatisation de certaines fonctions

va nécessairement restructurer le marché du travail dans les années à venir, en impactant notamment les métiers non-qualifiés, l'histoire des grands changements structurels montre que chaque emploi ayant été supprimé par la technologie, a été accompagné par la création de nouvelles activités. Une étude de McKinsey & Company conclue par exemple que, pour chaque emploi supprimé suite à l'évolution d'internet, 2.4 emplois ont été créés. Par ailleurs, il ne suffit pas de recruter des candidats intelligents pour qu'ils soient performants et engagés : c'est l'adéquation entre les exigences du poste, et les capacités du candidat, qui permet d'améliorer le rendement sur chaque poste et de limiter le turnover. Steffanie Wilk et Paul Sackett expliquent ainsi que, si le niveau de raisonnement d'un collaborateur dépasse les besoins du poste, celui-ci sera plus enclin à quitter (Wilk and Sackett, 2006).

L'utilisation d'un outil d'évaluation du raisonnement permet alors de recruter sur base d'un critère réellement explicatif de la réussite, mais aussi de connaitre précisément le degré de complexité qu'un candidat va pouvoir gérer, et quel type de tâches lui confier : des tâches simples et déjà expérimentées pour les personnes qui ont moins de facilités intellectuelles, ou des sujets nouveaux et stratégiques pour ceux avec la capacité de raisonner à des niveaux plus élevés. Pour répondre aux exigences des recruteurs et faciliter leur intégration dans les processus de recrutement, les tests de raisonnement ont ainsi fortement évolués, notamment grâce aux théories de l'évaluation basée sur le jeu. Ce nouveau format de tests cognitifs permet ainsi d'isoler une mesure fiable du niveau de raisonnement des candidats beaucoup plus rapidement, de renforcer l'engagement et la motivation des utilisateurs, ou encore de réduire leur niveau d'anxiété lors de la passation (Burgers, Eden, Van Engelenburg and Buningh, 2015). Aussi, pour rester pertinentes, les entreprises doivent revoir leurs critères de recrutement : à ce titre, le niveau de raisonnement d'un candidat constitue une donnée hautement prédictive de sa réussite future, et il est nécessaire de l'intégrer dans les démarches de recrutement, en complément d'analyses relatives à la personnalité et aux motivations. Ce changement de paradigme invite aussi à faire évoluer les systèmes d'éducation, en mettant l'accent sur la résolution de problèmes et l'analyse critique. Ces efforts sont nécessaires pour permettre à chacun d'exprimer ses talents et de trouver sa place dans la société de demain.

Lancé en 2020 par AssessFirst, BRAIN est un outil avant-gardiste qui répond à ce besoin, et qui est conçu pour mesurer le facteur *g*, ou l'intelligence générale, un indicateur clé de la performance professionnelle à travers une multitude de rôles. Cet outil se distingue par son adaptation aux tendances actuelles et aux exigences du marché, incarnées par trois principes essentiels :

- Une accessibilité optimale, permettant une utilisation fluide sur les appareils mobiles, répondant ainsi à la mobilité des utilisateurs modernes.
- Une approche adaptative, qui ajuste la difficulté des questions en temps réel selon les réponses du candidat, offrant une évaluation personnalisée et précise de ses capacités.
- Un design conçu pour engager et stimuler les candidats, rendant le processus d'évaluation non seulement plus agréable mais également plus immersif.

Ces innovations font de BRAIN un test à la pointe des innovations psychométriques, aligné avec les besoins évolutifs des clients et des candidats dans le contexte professionnel actuel. BRAIN optimise ainsi les conditions de passation, mais l'objectif reste le même : obtenir un indicateur précis, décontextualisé et transverse de la capacité de raisonnement d'un individu.

Le test BRAIN est une évaluation générale, conçue sans segmentation en diverses sections. Au cœur de cette expérience unique, le candidat est invité à résoudre une succession d'items logiques dans un temps défini. Ces items ne se résolvent pas par le biais de questions à choix multiples classiques, mais par la construction de réponses à partir d'éléments interactifs présents dans un environnement virtuel. Ce format créatif permet une évaluation plus dynamique des capacités cognitives. Aussi,

BRAIN repose sur des formats adaptatifs (CAT : Computerized Adaptive Testing). Ce système ajuste en continu le niveau de difficulté des items selon les performances du candidat : un premier item de difficulté intermédiaire est suivi par des items dont la difficulté varie en fonction des réponses précédentes, augmentant si la réponse est correcte, diminuant dans le cas contraire. Le test atteint son terme lorsque le candidat stabilise ses réponses à un niveau de difficulté constant, permettant ainsi une mesure précise de l'intelligence générale adaptée à chaque individu.

Le questionnaire BRAIN présente ainsi de nombreux avantages, parmi lesquels :

- **Une évaluation adaptative** : Les modèles adaptatifs (CAT : Computer Adaptive Testing) utilisés dans BRAIN permettent de créer un test sur-mesure pour chaque candidat, car les items qui lui sont présentés sont adaptés à son niveau et à sa performance réelle sur les items précédents. Cela permet ainsi : (1) de situer le niveau d'un candidat beaucoup plus rapidement, (2) d'améliorer l'expérience utilisateur car le candidat n'aura pas le sentiment d'être en situation d'échec, (3) de limiter la fuite des réponses et la triche, car aucun candidat n'a les mêmes items. Les modèles adaptatifs permettent ainsi de réduire le temps de passation, en proposant des items adaptés au niveau et aux performances réelles du candidat. En moyenne, un candidat passe entre 8 et 12 items sur la nouvelle version de BRAIN, ce qui correspond à une passation d'environ 10 minutes (contre 38 items et 21 minutes sur l'ancienne version).
- **Equitable, by design** : D'une part, BRAIN est pensé mobile-first. Cela permet de répondre à l'exigence grandissante des candidats de pouvoir postuler et compléter les tests de recrutement sur mobile. D'autre part, le matériel neutre utilisé dans BRAIN (c'est à dire un matériel exempt d'éléments verbaux et basé sur des couleurs sans lien avec la capacité de résolution des items), permet de rendre le test accessible aux personnes qui présentent un handicap comme la dyslexie.
- **Interculturelle** : Le matériel neutre utilisé par BRAIN permet de limiter les adaptations par langue.
- **Gamifiée** : BRAIN, sans être un jeu, fait appel aux mêmes leviers d'interactions et de motivations : feedback en temps réel, construction des réponses, niveau des items adapté, ou encore consignes personnalisées. Ce format de test permet de développer l'engagement des candidats tout en assurant la validité psychométrique de l'outil.
- **Solide** : Ree, Earles et Teachout (1994) démontrent que, pour prédire la performance au travail, la seule évaluation du facteur *g* est suffisante, et que l'ajout de compétences spécifiques (comme le raisonnement numérique ou verbal) n'apporte pas d'informations supplémentaires sur la capacité de réussite d'un candidat. Il est donc inutile de venir mesurer d'autres dimensions qui, même si elles sont intéressantes, allongent inutilement la longueur du test.

3. Bases théoriques

3.1. Éléments théoriques

BRAIN est un questionnaire psychométrique qui s'aligne avec les théories dominantes de l'intelligence en s'appuyant sur le concept du facteur *g*. S'inscrivant dans la lignée des travaux de Spearman (1904) et du modèle de Cattell, Horn et Carroll, ou CHC (1997), BRAIN mesure l'intelligence générale, reconnue comme le prédicteur le plus significatif de la performance professionnelle. La singularité de BRAIN réside dans sa focalisation exclusive sur le facteur g, s'écartant des versions antérieures du test qui différenciaient entre divers types de raisonnement comme le verbal, l'analogique, l'abstrait et le numérique. Ce recentrage est soutenu par des études, notamment celle de Ree, Earles et Teachout

(1994), qui ont démontré que le facteur g, à lui seul, offre une prédiction suffisante de la performance au travail, rendant les évaluations de compétences spécifiques redondantes. En effet, ces dernières ne semblaient pas fournir d'informations prédictives additionnelles significatives. Carroll (1993) appuie également cette approche en montrant que malgré l'existence de formes de raisonnement distinctes, ces dernières sont fortement intercorrélées et reliées au facteur *g*.

En outre, si l'analyse détaillée des différentes capacités cognitives peut sembler avantageuse, elle s'accompagne d'une augmentation considérable du temps de passation, qui n'est pas proportionnelle à la valeur prédictive ajoutée. Par conséquent, BRAIN a été conçu pour optimiser l'expérience des candidats, privilégiant la concision et l'efficacité. Cela aboutit à un questionnaire plus court, plus dynamique, qui respecte le temps des utilisateurs tout en leur fournissant une mesure fiable de l'intelligence générale. Ce choix méthodologique représente une avancée dans l'évaluation des talents, offrant aux entreprises un outil épuré et puissant pour leurs processus de sélection. Aussi, en concordance avec cette approche épurée, BRAIN s'insère dans une démarche plus globale de valorisation de l'efficience en matière d'évaluation des talents. En réduisant la durée du test, on respecte non seulement le temps des candidats, mais on favorise aussi une meilleure concentration et une performance optimale lors de la passation. En effet, comme peuvent le démontrer certaines études, quand les mesures de raisonnement sont trop longues, les résultats peuvent se voir influencés par la personnalité des répondants (Myszkowski, Storme, Kubiak and Baron, 2022). De plus, le questionnaire s'avère particulièrement pertinent dans un contexte où la rapidité et la pertinence de la prise de décision sont devenues primordiales. In fine, cette approche centrée sur le facteur *g* permet une intégration plus harmonieuse du test dans les systèmes de gestion des ressources humaines, qui peuvent ainsi bénéficier d'indicateurs plus synthétiques pour l'analyse et la prédiction des performances professionnelles. BRAIN se positionne donc comme un outil stratégique pour les entreprises soucieuses d'une évaluation cognitive efficace et concise.

3.2. Types d'analyses produits à partir de BRAIN

Facettes	Définition
Puissance de raisonnement	Il représente le facteur *g*, ou l'intelligence générale, sur une échelle de 1 à 10, sans décimale. Un score élevé suggère un fort potentiel d'apprentissage et une aptitude accrue à gérer des tâches complexes, indiquant ainsi un potentiel d'évolution professionnelle conséquent.
Vitesse de raisonnement	Mesurée par une jauge entourant le score, le temps indique la rapidité avec laquelle un individu complète le test. Une jauge complète signifie une utilisation maximale du temps imparti, reflétant une approche plus réfléchie ou prudente.
Tâches privilégiées	Cette dimension évalue le niveau de complexité des tâches qu'un individu est enclin à gérer, allant de simples actions routinières à des tâches intermédiaires autonomes, jusqu'à des activités complexes et stratégiques.
Type de prise de décision	Cette composante mesure le délai de prise de décision, classifiant les individus comme rapides, raisonnés ou prudents selon le temps qu'ils consacrent à la réflexion avant de se décider.
Style d'apprentissage	BRAIN identifie la manière dont une personne apprend le mieux, en déterminant si elle tend à innover, approfondir, expérimenter ou appliquer les connaissances et compétences.
Façon d'agir	Au-delà du niveau de raisonnement, BRAIN analyse la manière dont les réponses sont données, que ce soit dans les situations de réussite ou d'échec, offrant un aperçu du comportement sous-jacent.

Facettes	Définition
Points d'appuis	Dans la section "Points d'Appui" de BRAIN, trois éléments clés sont scrutés pour comprendre comment une personne aborde et traite l'information complexe, prend des décisions sous pression et maintient la précision dans ses jugements. La gestion de la complexité révèle sa capacité à appréhender des informations nuancées lors des prises de décision. La vitesse de décision indique si elle est encline à prendre des décisions rapidement ou si elle préfère un rythme plus mesuré. La précision, quant à elle, est un indicateur de la justesse et de la fidélité des décisions prises.
Facteurs de risques	BRAIN examine des "Facteurs de Risque" qui pourraient entraver la performance cognitive. La précipitation et la prudence excessive sont deux extrêmes qui reflètent respectivement une tendance à l'imprudence ou à l'hésitation excessive dans la prise de décision. Les déductions inexactes identifient une propension à commettre des erreurs d'analyse, tandis que l'indécision se manifeste par une réticence à s'engager fermement sur des réponses, souvent signalée par un recours fréquent à l'option "Je ne sais pas". Cet équilibrage entre les atouts et les risques potentiels offre une évaluation nuancée des capacités décisionnelles d'un individu.

Table 3.1. Descriptions des principales échelles mesurées par BRAIN.

4. Construction de BRAIN

La développement de BRAIN s'est déroulé selon plusieurs phases.

4.1. Phase 0

- **Étude de la littérature scientifique** : Cette étape vise à s'informer au maximum des évolutions scientifiques en termes d'évaluation des capacités de raisonnement en contexte professionnel, de nouveaux formats de questionnaires, d'impact de la gamification sur les formats d'évaluation, etc.
- **Développement et test d'un premier prototype** : Un premier prototype complet (consignes, items, etc.) a été développé. Celui-ci a par la suite été testé avec un échantillon de salariés AssessFirst. Ce premier test a permis d'avoir un premier retour sur la compréhension du format du questionnaire, la valeur de la consigne, le niveau des premiers items, la corrélation entre les scores à l'ancien BRAIN et la performance sur ce prototype, etc.
- **Amélioration du prototype** : Suite à cette première phase de test en interne, le prototype a été précisé, notamment en termes de consigne et de visualisation des éléments présents sur la scène de jeu. Un second test, avec un échantillon différent de salariés AssessFirst, a été conduit. Les modifications apportées sur le prototype ont été concluantes.

4.2. Phase 1

- **Mise en ligne de la première phase de test** : Une première version de test a été mise en ligne la première semaine d'octobre 2019. Cette version était composée de 20 items et visait surtout à avoir un premier retour des candidats sur le questionnaire, et à identifier les meilleurs items parmi les 20 proposés (les meilleurs items étant ceux qui apportent une information sur le niveau d'un candidat, qui permettent de discriminer les meilleurs des moins bons, et qui présentent une forte corrélation avec les scores de l'ancien BRAIN). Lors de cette phase, qui s'est déroulée entre la première semaine d'octobre 2019 et la deuxième semaine de novembre 2019, nous avons recueilli plus de 2000 passations. 6 items, considérés comme les « meilleurs » ont été isolés, et nous ont servi d'ancrage pour les phases de test suivantes.

4.3. Phase 2

- **Mise en ligne de la phase 2.1 de test** : Sur la base des analyses réalisées sur les données de la phase 1, 5 nouvelles séries de 20 items ont été créées. Dans chacune de ces séries ont été intégrés les 6 items d'ancrage isolés lors de la phase 1. Cette phase visait à analyser la qualité psychométrique des items conçus (est-ce qu'ils apportent une information sur le niveau d'un candidat, est-ce qu'ils permettent de discriminer les meilleurs des moins bons, et est-ce qu'ils présentent une forte corrélation avec les scores de l'ancien BRAIN). Lors de cette phase, qui s'est déroulée entre la deuxième semaine de novembre 2019 et la dernière semaine de novembre 2019, nous avons recueilli 150 passations par série, soit 750 passations au total.
- **Mise en ligne de la phase 2.2 de test** : Sur la base des analyses réalisées sur les données de la phase 2.1, 5 nouvelles séries de 20 items ont été crées. Dans chacune de ces séries ont été intégrés les 6 items d'ancrage isolés lors de la phase 1. Cette phase visait à analyser la qualité psychométrique des items conçus (est-ce qu'ils apportent une information sur le niveau d'un candidat, est-ce qu'ils permettent de discriminer les meilleurs des moins bons, et est-ce qu'ils présentent une forte corrélation avec les scores de l'ancien BRAIN). Lors de cette phase, qui s'est déroulée entre la dernière semaine de novembre 2019 et la dernière semaine de décembre 2019, nous avons recueilli 150 passations par série, soit 750 passations au total.
- **Mise en ligne de la phase 2.3 de test** : Sur la base des analyses réalisées sur les données de la phase 2.2, 5 nouvelles séries de 20 items ont été créées. Dans chacune de ces séries ont été intégrés les 6 items d'ancrage isolés lors de la phase 1. Cette phase visait à analyser la qualité psychométrique des items conçus (est-ce qu'ils apportent une information sur le niveau d'un candidat, est-ce qu'ils permettent de discriminer les meilleurs des moins bons, et est-ce qu'ils présentent une forte corrélation avec les scores de l'ancien BRAIN). Lors de cette phase, qui s'est déroulée entre la dernière semaine de décembre 2019 et la dernière semaine de janvier 2020, nous avons recueilli 150 passations par série, soit 750 passations au total.
- **Mise en ligne de la phase 2.4 de test :** Sur la base des analyses réalisées sur les données de la phase 2.3, 5 nouvelles séries de 20 items ont été créées. Dans chacune de ces séries ont été intégrés les 6 items d'ancrage isolés lors de la phase 1. Cette phase visait à analyser la qualité psychométrique des items conçus (est-ce qu'ils apportent une information sur le niveau d'un candidat, est-ce qu'ils permettent de discriminer les meilleurs des moins bons, et est-ce qu'ils présentent une forte corrélation avec les scores de l'ancien BRAIN). Lors de cette phase, qui s'est déroulée entre la dernière semaine de janvier 2020 et la deuxième semaine de février 2020, nous avons recueilli 150 passations par série, soit 750 passations au total.

4.4. Phase 3

- **Mise en ligne de la phase 3 de test** : Sur base des analyses réalisées sur l'ensemble des données de la phase 2, nous avons créé 5 nouvelles séries de 20 items. Chacune de ces séries est composée de : les 6 items d'ancrage identifiés lors de la phase 1 + 14 items issus de la phase 2 et considérés comme parmi les « meilleurs ». Au total, 70 items de la phase 2 de test ont ainsi été sélectionnés et intégrés à ces 5 nouvelles séries (donc 76 items au total, en comptant les 6 items d'ancrage). Cette phase visait à recueillir des données en vue de la calibration du questionnaire. Lors de cette phase, qui s'est déroulée entre la troisième semaine de février 2020 et la première semaine d'avril 2020, nous avons recueilli un peu plus de 500 passations par série, pour un total de 2800 passations.

4.5. Phase 4

- **Calibration et construction des modèles adaptatifs** : Les résultats de la phase 3 ont été analysés afin de réaliser les étalonnages pour les différents scores et indicateurs de BRAIN. Les modèles adaptatifs ont également été construits sur base des résultats et niveaux des items de la phase 3. Cette phase de calibration s'est déroulée entre la deuxième semaine d'avril 2020 et la dernière semaine d'avril 2020.
- **Migration** : Le test adaptatif a ensuite été intégré à l'application (en simulation et non sur la production) à la place de l'ancien BRAIN. Les correctifs nécessaires à son intégration (modifications des rapports, modification de certains visuels, migration des anciens scores, etc.) ont été réalisés lors de cette phase. Elle s'est déroulée entre la deuxième semaine d'avril 2020 et la deuxième semaine de mai 2020..
- **Test QA** : La troisième semaine de mai a été consacrée aux tests QA
- **Mise en production :** Mise en production le mardi 9 juin 2020.

La méthodologie déployée pour le développement de BRAIN reflète un processus de recherche et développement rigoureux et itératif, caractérisé par une série de six phases de tests et de collecte de données. Chaque phase était méthodiquement conçue pour affiner le test en fonction des retours, des performances des items, et de leur corrélation avec les versions antérieures, garantissant ainsi une évolution fondée sur des preuves tangibles. L'ampleur de ce projet est illustrée par le nombre de passations, environ 8000, qui ont joué un rôle déterminant dans le passage de la version initiale à la version finale. Ce processus exhaustif, étalé sur plusieurs mois, a permis d'aboutir à un outil BRAIN calibré avec précision, doté de modèles adaptatifs sophistiqués et prêt pour une intégration réussie dans l'application finale. La mise en œuvre de cette méthodologie démontre un engagement envers l'excellence scientifique et l'expérience utilisateur, affirmant BRAIN comme un instrument d'évaluation de l'intelligence générale de pointe pour le milieu professionnel.

5. Validité

Comment savoir si un questionnaire mesure réellement ce qu'il est censé mesuré ? Comment s'assurer de la bonne mesure de chaque échelle ainsi que la juste signification des résultats au questionnaire ? Les réponses à ces questions sont obtenues grâce à sa validation. L'objectif de la validation d'un questionnaire consiste à confirmer que celui-ci mesure réellement ce qu'il cherche à mesurer, et à quel point les résultats que nous

pouvons en tirer son précis. Historiquement, la validité a été définie comme une corrélation entre un score à un questionnaire et un critère externe, qui mesure soit le même construit, ou qui mesure un construit censé être en lien avec le construit associé à ce score. Plusieurs types de validité sont nécessaires pour établir et assurer la validité d'un questionnaire. Les études de validité de BRAIN portent sur les types de validité suivants :

- **Validité de construit** : elle se réfère à la mesure dans laquelle le questionnaire évalue réellement la dimension ou le construit psychologique qu'il est censé évaluer. Sont notamment proposées des analyses relatives au RMSEA et aux paramètres de distribution ;
- **Validité convergente** : la validité convergente fait référence au degré auquel deux mesures de construits qui devraient théoriquement être liés le sont. En d'autres termes, la validité convergente mesure à quel degré les résultats d'un questionnaire sont corrélés avec ceux d'un autre questionnaire qui évalue le même concept ;
- **Validité prédictive** : la validité prédictive d'un questionnaire de raisonnement est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire de raisonnement peuvent être utilisés pour prédire la performance future.

5.1. Validité de construit

La validité de construit permet de savoir si le test mesure réellement le construit théorique souhaité, ou autre chose. Ce type de validité chevauche certains des autres aspects de la validité, car toute preuve de validité contribue à la compréhension de la validité de construit d'un instrument d'évaluation. La validité de construit est importante, car elle influe sur l'interprétation des scores d'un questionnaire. Si un questionnaire prétend mesurer une dimension spécifique, comment être sûr qu'il mesure réellement cette dimension ? Si un questionnaire est censé mesurer une dimension spécifique, mais qu'en réalité, ce n'est pas le cas, toute interprétation du score est incorrecte et pourrait conduire à des décisions biaisées. La validité de construit ne cherche pas simplement à savoir si un questionnaire mesure une dimension. Aussi, elle considère des investigations complexes cherchant à savoir si les interprétations des résultats des tests sont cohérentes avec un réseau impliquant des termes théoriques et d'observation (Cronbach & Meehl, 1955).

Il n'existe pas de méthode unique pour déterminer la validité de construit, mais différentes méthodes et approches sont combinées. Pour évaluer la validité de construit de BRAIN, nous privilégions ici deux méthodes complémentaires, à savoir le calcul du RMSEA et les paramètres de distribution.

- **le RMSEA**, ou Root Mean Square Error of Approximation, est une mesure d'ajustement qui est souvent utilisée pour évaluer l'adéquation d'un modèle de mesure. Le RMSEA mesure la différence entre les données observées et les données ajustées du modèle, corrigée pour le nombre de paramètres libres du modèle. Plus précisément, le RMSEA évalue l'ajustement absolu du modèle en comparant la variance non expliquée dans les données avec la variance non expliquée attendue dans les données selon le modèle. Généralement, un RMSEA < .05 indique un bon ajustement du modèle aux données (Steiger & Lind, 1980; Browne & Cudeck, 1993) ;
- **les paramètres de distribution** correspondent à la distribution des scores dans les questionnaires. Ils permettent d'identifier les scores atypiques, d'explorer les différences individuelles dans la distribution des scores, et de mieux interpréter les résultats.

5.1.1. RMSEA

Les dernières études de RMSEA pour BRAIN ont été conduites en 2020. Pour le score global au test, l'indice RMSEA est inférieur à .05 (RMSEA = .04), indiquant une bonne adéquation du modèle aux données. Cela suggère que le modèle a une bonne capacité à expliquer les relations entre les variables mesurées et que les différences entre les données observées et les données prédites par le modèle sont faibles. Ces informations apportent une preuve de validité de construit supplémentaire au questionnaire BRAIN.

Dimensions	*RMSEA*
Global score	.04

Table 5.1. RMSEA de l'échelle principale de BRAIN.

5.1.2. Paramètres de distribution

La distribution des scores obtenus par un questionnaire est un aspect important de la validité de construit du questionnaire. En effet, la manière dont les scores sont distribués pour chaque dimension de motivation peut fournir des informations importantes sur la manière dont le test mesure réellement cette dimension, ainsi que sur la manière dont les scores sont interprétés. Nos analyses de paramètres de distribution se centrent sur 5 paramètres principaux et nécessaires :

- la **moyenne**, qui est un indicateur de la tendance centrale des scores dans la distribution ;
- la **médiane**, qui est la valeur qui divise la distribution en deux parties égales : la moitié des scores sont supérieurs à la médiane, et l'autre moitié sont inférieurs. C'est également un indicateur de la tendance centrale des scores dans la distribution, qui est moins sensible aux scores extrêmes que la moyenne ;
- l'**écart-type**, qui est une mesure de la dispersion des scores autour de la moyenne. Il est calculé en prenant la racine carrée de la variance des scores ;
- l'**asymétrie** , qui est calculée en comparant la fréquence des scores à gauche et à droite de la moyenne. Si la distribution est parfaitement symétrique, l'asymétrie est nulle. Si la distribution est asymétrique vers la gauche, l'asymétrie est négative. Si la distribution est asymétrique vers la droite, l'asymétrie est positive ;
- l'**aplatissement** (ou coefficient d'aplatissement), qui est calculé en comparant la distribution des scores à une distribution normale. Si la distribution est moins aplatie que la distribution normale, l'aplatissement est négatif. Si la distribution est plus aplatie que la distribution normale, l'aplatissement est positif.

Les attentes pour les paramètres de distribution dépendent du contexte et de l'instrument de mesure utilisé. Toutefois, en général, voici ce que l'on attend pour de « bons » paramètres de distribution :

- La moyenne doit être proche de la valeur médiane : cela indique que la distribution est symétrique. Si la moyenne est significativement différente de la médiane, cela peut indiquer une asymétrie de distribution ;
- L'écart-type doit être raisonnable et suffisamment grand pour capturer les différences individuelles dans la dimension mesurée, mais pas trop grand pour ne pas diluer les différences entre les individus. En général, on s'attend à ce que l'écart-type soit d'environ 2 pour les échelles de personnalité en 10 points ;
- L'asymétrie (skewness) doit être proche de 0 (distribution symétrique). Si l'asymétrie est significativement différente de 0, cela peut indiquer une asymétrie dans la distribution ;
- L'aplatissement (kurtosis) doit être proche de 0 (distribution normale). Si l'aplatissement est significativement différent de 0, c'est que la distribution est soit plus plate, soit plus pointue.

Les dernières études de paramètres de distribution de BRAIN ont été conduites en 2023 (N = 50000).

Dimension	Moyenne	Médiane	Écart-type	Asymétrie	Aplatissement
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. Paramètres de distribution du score et du temps global.

Les paramètres de distribution sont ainsi cohérents avec les standards attendus. La normalité des distributions peut être interprétée par des indices comme la superposition des moyennes et des médianes, et à l'aide des coefficients de symétrie et d'aplatissement qui sont proches de 0.

5.2. Validité convergente

5.2.1. Introduction

La validité convergente est une mesure de la similitude entre les scores d'un test de raisonnement et ceux d'autres tests ou mesures qui évaluent la même dimension ou facteur de raisonnement. En d'autres termes, elle permet de vérifier si un test mesure effectivement ce qu'il est censé mesurer. Plus spécifiquement, la validité convergente s'intéresse à la corrélation entre les scores d'un test et ceux d'autres mesures ou tests qui évaluent les mêmes dimensions de raisonnement. Une forte corrélation entre les scores indique que les échelles sont convergentes et mesurent le même construit, ce qui renforce la validité.

Il convient de noter qu'il n'existe pas de seuil « officiel » pour juger de la qualité de la convergence entre les deux mesures. Aussi, le seuil approprié dépend du contexte spécifique dans lequel le questionnaire est utilisé et des caractéristiques de la population cible. De plus, la validité convergente doit être évaluée en conjonction avec d'autres mesures de validité pour avoir une évaluation complète de la qualité du test de personnalité. Cependant, plusieurs auteurs et recherches ont proposé quelques suggestions ou seuils de satisfaction : (1) une corrélation de .7 ou plus entre un questionnaire et d'autres mesures qui évaluent la même dimension est un indicateur d'une très forte validité convergente par Campbell et Fiske (1959), (2) une corrélation de .6 est recommandée comme seuil de validité par Worthington et Whittaker (2006), (3) une corrélation de .5 ou plus est considérée comment une bonne validité convergente par Bagozzi et Yi (1988) et par Revelle et Condon (2015), (4) une corrélation de .4 est considérée comme acceptable par Nunnally et Bernstein (1994). En somme, s'il n'existe pas de consensus clair ou de règle d'or (Marsh, Hau & Wen, 2004) sur la valeur exacte à utiliser comme seuil de validité convergente, il est recommandé de viser des corrélations de .5 minimum pour justifier d'une bonne validité d'un questionnaire de raisonnement.

5.2.2. Validité convergente avec IMak-17

Dans le cadre du développement de BRAIN, nous étudions la validité convergente du questionnaire, avec un test des matrices généré spécifiquement dans le cadre de notre étude. Il s'agit dun test psychométrique inspiré des matrices de Raven qui est utilisé pour mesurer l'intelligence fluide, c'est-à-dire la capacité de raisonner et de résoudre de nouveaux problèmes, sans faire appel à des connaissances antérieures. Le test est composé d'une série de matrices ou de motifs visuels où le participant doit identifier la partie manquante pour compléter la matrice. Il est souvent utilisé dans les contextes éducatifs et professionnels pour évaluer la capacité de réflexion logique et d'analyse visuelle. Cette étude a été réalisée en 2021.

5.2.2.1. Participants et procédure

Pour cette étude, nous avons recruté des participants inscrits sur l'application AssessFirst. Un lien vers l'étude a été envoyé par e-mail aux 11 000 derniers utilisateurs de l'application, qui avaient auparavant complété BRAIN, et qui avaient indiqué leur consentement à être contactés pour des études scientifiques. L'e-mail précisait que cette enquête était liée au test BRAIN qu'ils avaient déjà passé sur la plateforme, mais que leurs réponses et résultats seraient utilisés uniquement à des fins scientifiques, sans impact sur leur profil d'évaluation en ligne et sans communication à des entreprises externes. Pour cette raison, cette étude représente un contexte à faible enjeu. 7.4 % des utilisateurs ont répondu volontairement. Aussi, le test des matrices n'étant pas optimisé pour un affichage sur les écrans de smartphone, nous avons exclu les participants qui avaient complété l'enquête sur un smartphone et n'avons retenu que ceux qui l'avaient complété sur un ordinateur. Dans cet échantillon final (N=555), 53.5 % des participants se sont identifiés comme femmes et 46.5 % comme hommes, avec un âge moyen de 38.9 ans. Toutes les conditions et exclusions de données sont rapportées dans ce document. Les participants ont été contactés sur la base de leur réponse antérieure au test BRAIN, mais certains d'entre eux peuvent avoir pris d'autres mesures sur la plateforme AssessFirst ; cependant ces mesures n'ont pas été utilisées dans le cadre de cette étude.

Une fois connectés, les participants ont effectué le test de matrices, composé de 17 items. Le temps passé à répondre à chaque item du test de matrices a été enregistré, ainsi que l'option de réponse sélectionnée. Les temps de réponse ont été enregistrés directement via l'application d'enquête utilisée (SurveyGizmo), en calculant la différence (en secondes) entre le moment de l'affichage de la page (c'est-à-dire l'item) et le moment où le participant a cliqué pour soumettre sa réponse. À la fin de l'étude, les réponses des participants ont été reliées aux données de l'application AssessFirst, ce qui a permis d'utiliser le questionnaire BRAIN précédemment passé, et d'étudier la validité convergente entre les deux tests.

5.2.2.2. Mesure

Nous avons utilisé la bibliothèque IMak (Blum & Holling, 2018) pour générer des analogies sous forme de matrices afin d'évaluer l'intelligence. Le test final se composait de 4 items d'entraînement à une règle, 4 items à une règle, 6 items à deux règles, 6 items à trois règles et 1 item à quatre règles, présentés dans cet ordre. Dans le cadre de cette étude, ce test sera désigné par le terme IMak-17. Étant donné que le nombre de règles est un indicateur de la difficulté de l'item (Blum & Holling, 2018), ce test peut être considéré comme un test de matrices progressif. Nous avons estimé sa fidélité en utilisant l'omega de McDonald (Flora, 2020), qui était de $\omega = 0.84$.

5.2.2.3. Objet de l'étude

Dans le cadre de notre étude, nous avons entrepris de tester la validité convergente de BRAIN en le comparant à l'IMak-17. Pour rappel, la validité convergente est un type de validité qui mesure dans quelle proportion deux tests indépendants, censés évaluer le même construit psychologique, sont en corrélation. Pour ce faire, nous utilisons une analyse de corrélation statistique, qui évalue le degré de relation linéaire entre les scores obtenus sur les deux tests. Si nous observons une corrélation forte entre BRAIN et l'IMak-17, cela indiquerait que les deux mesurent des aspects similaires de l'intelligence, suggérant ainsi une forte validité convergente. En revanche, une corrélation faible suggérerait que les tests pourraient ne pas mesurer le même construit ou que l'un des tests pourrait ne pas être un bon indicateur de l'intelligence.

La force de la corrélation est généralement mesurée à l'aide du coefficient de corrélation de Pearson (r). Les valeurs de ce coefficient varient de -1 à +1, où +1 indique une corrélation positive parfaite, -1 une corrélation négative parfaite, et 0 aucune corrélation. En règle générale, une valeur de r au-delà de 0,70 est considérée comme forte, tandis que des valeurs autour de 0,50 sont considérées comme modérées et celles en dessous de 0,30 sont faibles. Pour notre analyse, nous avons collecté les scores des participants sur BRAIN et les comparons avec leurs scores sur l'IMak-17. Nous calculons ensuite le coefficient de

corrélation de Pearson pour évaluer la force de la relation entre les deux ensembles de scores. Si une forte corrélation est trouvée, cela nous permettra de conclure avec plus de confiance que le BRAIN possède une bonne validité convergente en ce qui concerne la mesure de l'intelligence, similaire à celle de l'IMak-17.

Il est important de noter que les résultats peuvent être influencés par le contexte de passation. En effet, la corrélation pourrait probablement être atténuée par le temps écoulé entre les deux mesures, les différences dans le contenu des items (surtout en raison de l'aspect adaptatif et gamifié du test BRAIN, basé sur les CAT), et, plus important encore, le fait que l'IMak-17 ait été complété par les participants dans des conditions à faible enjeu, alors que le test BRAIN a été complété dans des conditions à enjeu élevé (car les résultats au test BRAIN sont utilisés pour alimenter le profil en ligne de la personne dans l'application).

5.2.2.4. Résultats

Les analyses font état d'un coefficient de corrélation de Pearson r = 0.74. Cette corrélation de 0.74 est significative et suggère une forte relation positive entre les scores obtenus sur BRAIN et ceux obtenus sur l'IMak-17. En termes de validité convergente, cela indique que BRAIN est un indicateur robuste de l'intelligence, similaire à l'IMak-17. Un coefficient de cette ampleur nous permet de conclure que les deux tests évaluent, avec un degré élevé de similitude, le même construit psychologique, c'est-à-dire l'intelligence.

Aussi, une corrélation aussi élevée suggère également que les variations dans les performances des participants sur l'un des tests peuvent être prédites de manière fiable par leurs performances sur l'autre test. Cela renforce la crédibilité de BRAIN en tant qu'outil d'évaluation cognitive et justifie son utilisation dans des contextes où une mesure précise de l'intelligence est nécessaire. Il est important de noter que bien que la corrélation soit forte, elle n'est pas parfaite. Cela indique que, bien que les tests partagent une variance commune significative, il y a aussi des différences qui pourraient être attribuées aux spécificités de chaque test ou à d'autres facteurs mesurant des aspects légèrement différents de l'intelligence. Néanmoins, une corrélation de 0.74 est considérée comme une preuve solide de validité convergente.

5.3. Validité prédictive

La validité prédictive d'un questionnaire de raisonnement est la mesure de sa capacité à prédire une variable cible, comme la performance ou le turnover. En d'autres termes, il s'agit de savoir si les résultats au questionnaire peuvent être utilisés pour prédire la performance future. La preuve de validité prédictive est particulièrement applicable lorsque l'on souhaite faire une inférence, à partir d'un score du questionnaire, de la position de l'évalué sur une autre variable de critère évaluée de manière indépendante à une date ultérieure. Les études de validité prédictive liées à BRAIN sont présentées dans un guide dédié.

5.4. Conclusion

La validation d'un questionnaire de raisonnement est cruciale pour s'assurer que les mesures obtenues sont précises. Dans cette étude, nous avons examiné la validité de construit, la validité convergente, et la validité prédictive de BRAIN. Nos analyses montrent que : (1) le questionnaire est bien structuré et montre une bonne homogénéité de la mesure, (2) BRAIN présente une forte validité convergente avec l'IMak-17, (3) BRAIN est prédictif de la réussite en poste. Dans l'ensemble, les résultats obtenus répondent aux standards psychométriques les plus exigeants, et démontrent la validité de BRAIN. Aller plus loin dans les qualités psychométriques de BRAIN requiert toutefois de s'intéresser à sa fidélité. En ce sens, un questionnaire doit être valide et fidèle pour être utilisé dans le cadre de décisions professionnelles (recrutement, mobilité, etc.). En effet, un questionnaire valide mais non fidèle signifierait que le test mesure bien ce qu'il doit mesurer, mais que les scores individuels sont incohérents. Au contraire, un questionnaire valide et fidèle signifie que le questionnaire mesure systématiquement ce qu'il est censé mesurer : en d'autres mots, il frappe constamment dans le mille. Les preuves de fidélité sont présentées dans le chapitre suivant.

6. Fidélité

Comment savoir si les résultats d'un questionnaire sont fiables ? Comment s'assurer que le questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ? Les réponses à ces questions sont obtenues grâce à l'étude de sa fidélité. Alors que la validité renseigne sur la capacité d'un questionnaire à réellement mesurer ce que l'on souhaite mesurer, la fidélité permet de savoir si cette mesure sera consistante et fiable à chaque fois que ce même questionnaire est complété par une même personne. En somme, la fidélité d'un questionnaire est une mesure de sa cohérence ou de sa stabilité dans le temps. Elle vise à déterminer si un questionnaire produit des résultats similaires lorsque les mêmes questions sont posées à une même personne à différents moments ou à des personnes similaires. L'objectif de la fidélité est donc de garantir que les résultats obtenus sont fiables et précis. La fidélité de BRAIN est ici analysée à travers la fidélité test-retest. Il s'agit d'une méthode pour évaluer la fidélité d'une mesure en mesurant la même variable à deux moments différents. Elle permet de mesurer la stabilité temporelle de la mesure et d'estimer la proportion de la variance totale qui est attribuable à l'erreur de mesure. Elle est souvent utilisée dans les études longitudinales ou pour évaluer la stabilité d'un test sur une période de temps donnée. Elle consiste à administrer le même questionnaire à un groupe de participants à deux moments différents, avec un intervalle de temps entre les deux passations. La corrélation entre les résultats est calculée pour déterminer la fidélité. Si celle-ci est élevée, cela signifie que les scores sont stables et que le test peut donc être considéré comme fiable.

Les dernières études de fidélité test-retest pour BRAIN ont été conduites en 2023 (N = 2309). Les deux passations ont été administrées avec un intervalle de 3 mois au minimum. Le score global au questionnaire BRAIN (capacité générale de raisonnement) démontre une bonne fidélité test-retest avec un coefficient r = .81. Cette stabilité des scores sur une période significative souligne la capacité de BRAIN à fournir une mesures cohérente et fiable du niveau de raisonnement d'une personne, un atout essentiel pour les évaluations en contexte professionnel où les décisions à long terme se basent sur ces données.

Dimensions	*Pearson r*
Global score	.81

Table 6.1. Fidélité test-retest du score global au BRAIN.

7. Sensibilité

La sensibilité, aussi appelée discrimination, désigne la capacité d'un questionnaire à distinguer les personnes ayant un niveau élevé sur une dimension et les personnes ayant un niveau faible. Elle reflète donc la capacité du questionnaire à identifier la singularité de chaque individu. Un questionnaire de raisonnement sensible peut ainsi identifier les différences subtiles entre les personnes, et aider à comprendre leurs comportements avec davantage de précision et de discrimination. La sensibilité se réfère ainsi à la capacité du questionnaire à identifier correctement les personnes qui possèdent la caractéristique mesurée et à éviter les faux positifs (personnes qui ne possèdent pas la caractéristique, mais qui sont identifiées comme telles par le questionnaire). Cette propriété est directement liée à la qualité du score.

Dans le cadre du développement de BRAIN, nous étudions le pouvoir discriminant et la sensibilité de chaque item du test grâce à la mesure du paramètre alpha α. Ce paramètre de discrimination fait partie intégrante des modèles de la théorie de réponse à l'item (IRT), qui est une approche moderne de la conception des tests psychométriques. Le paramètre de discrimination, α, reflète la capacité d'un item à différencier entre les répondants selon leur niveau de la capacité ou du trait mesuré. Un item avec un haut

paramètre de discrimination sera plus efficace pour distinguer les individus qui ont des niveaux différents de la dimension en question. Par exemple, dans un test de compétence mathématique, un item avec une discrimination élevée permettra de mieux différencier entre les individus ayant une compétence mathématique élevée et ceux ayant une compétence faible. En pratique, les paramètres de discrimination sont estimés lors du calibrage des items et peuvent être utilisés pour sélectionner les items les plus informatifs pour un test. Ils jouent notamment un rôle dans la construction de tests adaptatifs (CAT), où les items sont choisies dynamiquement pour correspondre au niveau de compétence estimé du répondant.

Le paramètre α est un indicateur crucial dans l'évaluation de la discrimination d'un item dans les tests psychométriques, servant à mesurer sa capacité à différencier les individus en fonction de leur niveau d'aptitude. Selon la théorie de réponse à l'item, la valeur du paramètre α a une relation directe avec le pouvoir discriminant de l'item : plus la valeur de α est élevée, meilleure est la discrimination. Pour faciliter l'interprétation des valeurs de α, Baker (2001) a introduit une grille d'évaluation normative qui catégorise la capacité discriminante des items comme suit :

- La discrimination est considérée comme « nulle » si α égale 0, signifiant que l'item ne différencie pas entre les niveaux de capacité des répondants.
- Elle est jugée « très faible » si α est compris entre 0,01 et 0,34, indiquant que l'item a une capacité minimale de discrimination.
- La discrimination est « faible » pour des valeurs de α allant de 0,35 à 0,64, montrant une capacité limitée à distinguer les répondants selon leur aptitude.
- Une valeur de α entre 0,65 et 1,34 correspond à une « bonne » discrimination, indiquant une bonne efficacité pour la séparation des niveaux de capacité.
- La discrimination est qualifiée de « forte » si α se situe dans l'intervalle de 1,35 à 1,69, dénotant une capacité supérieure à différencier les individus efficacement.
- Enfin, la discrimination est considérée comme « très forte » lorsque α dépasse 1,70, signifiant que l'item est extrêmement efficace pour identifier les différences de capacité parmi les répondants.

Cette grille est un outil précieux lors de la conception et de l'évaluation de tests, car elle fournit des repères clairs pour juger de la qualité et de l'utilité des items au sein d'une échelle de mesure.

Les dernières études sensibilité de BRAIN, basées sur le calcul du paramètre α, ont été conduites en 2020. L'évaluation du pouvoir de discrimination des items du test BRAIN révèle une forte capacité à distinguer les individus en fonction de leur aptitude. Cette évaluation s'appuie sur la distribution des valeurs de discrimination des items individuels, classées selon la grille d'évaluation de Baker (2001) :

- Il est encourageant de noter qu'aucun des items de BRAIN ne présente une discrimination « nulle », ce qui signifie qu'il n'y a pas d'items incapables de différencier les répondants selon leur aptitude.
- De même, aucun item ne se situe dans la catégorie de discrimination « très faible ». Cela indique l'absence d'items ayant seulement une capacité marginale à distinguer les niveaux de capacité.
- Une petite portion de l'inventaire, soit 6 items, affiche une discrimination « faible ». Bien que ces items contribuent moins efficacement à la différenciation du niveau d'aptitude, ils peuvent toujours fournir des informations utiles dans le contexte global du test.
- La majorité des items, avec un total de 31, se classent dans la catégorie de « bonne » discrimination. Ces items sont capables de bien distinguer entre les individus à différents niveaux de capacité et constituent l'épine dorsale solide du test.

- 16 items ont été identifiés avec une discrimination « forte », ce qui témoigne de leur excellente capacité à différencier avec précision entre les personnes ayant des niveaux de trait variés.
- Enfin, un nombre notable de 23 items démontre une discrimination « très forte ». Ces items sont particulièrement puissants et efficaces pour séparer les répondants selon leur niveau d'aptitude, contribuant ainsi de manière significative à la sensibilité globale du test.

En somme, le test BRAIN se caractérise par une qualité psychométrique élevée, avec une prédominance d'items offrant une discrimination allant de bonne à très forte. La présence d'items à discrimination faible dans le test BRAIN est une décision délibérée qui enrichit la dimensionnalité du test et sa capacité à évaluer l'ensemble du spectre de capacités. Ces items, bien que moins discriminants dans le contexte général, sont précieux car ils fournissent des informations spécifiques sur les extrémités du spectre de difficulté. Les items à faible discrimination sont particulièrement utiles pour évaluer les individus aux capacités très faibles ou très élevées. Pour les répondants aux compétences très faibles, ces items peuvent être parmi les seuls qu'ils peuvent répondre correctement, offrant ainsi une mesure de capacité là où les items plus discriminants ne le pourraient pas. À l'autre extrémité, pour les individus aux capacités très élevées, ces items peuvent aider à identifier le seuil auquel ils commencent à rencontrer des difficultés. Ainsi, même si ces items à faible discrimination contribuent moins à la différenciation des répondants au milieu du spectre de capacités, ils sont essentiels pour capturer la variabilité aux deux extrémités. Cela permet de maintenir un test équilibré, qui peut fournir des mesures pertinentes pour tous les niveaux de compétence, assurant ainsi une évaluation complète et nuancée de l'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. Sensibilité des items BRAIN.

Les statistiques descriptives concernant le pouvoir de discrimination des items du test BRAIN mettent en lumière une tendance générale vers une discrimination élevée. Les valeurs moyenne et médiane, très proches l'une de l'autre, se situent respectivement à 1,40 et 1,39, indiquant une discrimination forte selon la classification de Baker (2001). Cette proximité entre la moyenne et la médiane suggère également une distribution relativement symétrique des valeurs de discrimination autour d'un niveau élevé, ce qui est caractéristique d'un test bien calibré. Le minimum observé, de 0,44, se trouve dans la catégorie de discrimination faible. Bien que ce soit la plus basse valeur de discrimination parmi les items du test, elle n'atteint pas le seuil de discrimination très faible ou nulle, ce qui signifie que même l'item le moins discriminant apporte encore une certaine valeur au test. À l'opposé, le maximum relevé de 3,60 est bien au-delà du seuil de discrimination très forte. Cette valeur extrême reflète un item particulièrement puissant qui peut très efficacement distinguer les participants selon leurs niveaux de capacité. Ces résultats montrent que le test BRAIN, dans son ensemble, possède une forte capacité discriminatoire, avec une distribution des valeurs de discrimination qui favorise une évaluation précise des capacités cognitives sur une vaste gamme. L'existence d'items ayant des valeurs de discrimination à la fois faibles et très élevées permet au test de couvrir efficacement l'ensemble du spectre de difficulté, garantissant ainsi une mesure complète de l'intelligence à travers divers niveaux de compétence.

8. Équité

L'équité dans le contexte d'un questionnaire de raisonnement fait référence à la mesure dans laquelle celui-ci est équitable et impartial pour toutes les personnes qui le passent, quelle que soit leur origine, leur sexe, leur orientation sexuelle, leur race ou leur culture. En d'autres termes, un questionnaire équitable est conçu pour être objectif et juste pour toutes les personnes qui le passent, sans biais ou discrimination à l'encontre d'un groupe particulier. Nos équipes mettent en ce sens tout en œuvre pour veiller au caractère équitable des questionnaires et des analyses prédictives, et pour s'assurer que l'usage de nos algorithmes dans les processus décisionnels n'amènent pas de discriminations via des biais algorithmiques insoupçonnés.

8.1. Équité dans les résultats

Les données présentées dans cette section démontrent qu'il n'existe pas de différences majeures ou de taille d'effet forte dans les résultats au questionnaire BRAIN en fonction des variables de genre et d'âge. Note : seules les informations personnelles nécessaires à l'utilisation correcte d'AssessFirst sont demandées à nos utilisateurs. Par exemple, il n'y a aucune mention de l'orientation religieuse, politique ou sexuelle, quel que soit le moment. Quand il s'agit de l'âge, nous demandons la date de naissance afin de nous assurer que cela n'affecte pas la façon dont les questions sont traitées. De même, toutes les variables ci-après analysées n'interviennent à aucun moment dans le calcul des résultats dans la solution.

8.1.1. Équité selon le genre

Les dernières analyses d'équité selon le genre pour BRAIN ont été conduites en 2022, sur un échantillon (N = 332587) composé de 51% d'hommes et 49% de femmes. Globalement, les moyennes du score global au BRAIN se situent toutes les deux entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Il n'existe donc pas de différence majeure entre les résultats des hommes et ceux des femmes sur le score global mesuré par le questionnaire BRAIN. Les résultats sont ainsi équitable selon le genre du répondant.

Dimension	Hommes	Femmes
Global score	5.74	5.59

Table 8.1. Moyennes du score global en fonction du genre.

Aussi, l'effet ici mis en lumière vient supporter la conclusion précédente. En effet, le d de Cohen est égal à .07, ce qui démontre l'absence de différence entre les résultats des hommes et ceux des femmes.

Dimension	d de Cohen	Taille d'effet
Global score	.07	-

Table 8.2. d de Cohen pour le genre du score global.

8.1.2. Équité selon l'âge

Les dernières analyses d'équité selon le genre pour BRAIN ont été conduites en 2022, sur un échantillon (N = 64310) composé de 52% de personnes de plus de 35 ans et 48% de personnes de moins de 35 ans. Globalement, les moyennes du score global au BRAIN se situent toutes les deux entre 5 et 6, ce qui est proche de la moyenne théorique de 5.5. Il n'existe donc pas de différence majeure entre les résultats des personnes de plus de 35 ans et celles de moins de 35 ans sur le score global mesuré par le questionnaire BRAIN. Les résultats sont ainsi équitable selon l'âge du répondant.

Dimension	≥ 35 ans	< 35 ans
Global score	5.45	5.67

Table 8.3. Moyennes du score global en fonction de l'âge.

Aussi, l'effet ici mis en lumière vient supporter la conclusion précédente. En effet, le d de Cohen est égal à -.11, ce qui démontre l'absence de différence entre les résultats des personnes de plus de 35 ans et celles de moins de 35 ans sur le score global mesuré par le questionnaire BRAIN.

Dimension	d de Cohen	Taille d'effet
Global score	-.11	-

Table 8.4. d de Cohen pour l'âge du score global.

8.1.3. Conclusion

Que cela soit en termes de genre ou de catégories d'âge, il n'existe pas de différences majeures entre les résultats obtenus par les différents groupes au questionnaire BRAIN d'AssessFirst. En somme, les résultats obtenus permettent de soutenir l'hypothèse selon laquelle le questionnaire proposé ne discrimine aucun public et s'avère équitable.

Conclusion

Les études psychométriques présentées dans ce manuel technique démontrent et attestent de la solidité scientifique des questionnaires développés par AssessFirst. En ce sens, les différentes analyses proposées montrent la validité, la fidélité, la sensibilité et l'équité de chaque questionnaire. Il est important de souligner que ces résultats ont été obtenus grâce à un processus rigoureux de développement et de validation des questionnaires, qui respectent les normes internationales les plus strictes en matière de psychométrie. La mise en conformité de ces outils avec les standards préconisés par l'Association Américaine de Psychologie (A.P.A.) et la Commission Internationale des Tests (I.T.C.) permet aujourd'hui à AssessFirst de garantir un haut niveau de qualité dans la conception des questionnaires et d'améliorer continuellement la fidélité de ses outils d'évaluation. Ces efforts et ce devoir de qualité permettent de répondre aux exigences des professionnels des ressources humaines en matière d'évaluation des candidats et des collaborateurs.

D'autres analyses seront régulièrement ajoutées à ce manuel afin de parfaire cette démonstration de robustesse scientifique. Aussi, la roadmap des prochaines études inclut : (1) la validité prédictive de SWIPE - en fin d'année 2023, (2) la fidélité test-retest de SWIPE en janvier 2024, (3) des études liées à l'équité de niveau hiérarchique de SWIPE.

Pour plus d'informations sur les aspects scientifiques relatifs à nos outils et à notre produit, vous pouvez contacter votre Account Manager et/ou Customer Success référent, où l'un de nos experts ci-dessous.

Emeric KUBIAK
Psychologue
Head of Science @AssessFirst

Simon BARON
Psychologue
Chief Product Officer @AssessFirst

Bibliographie

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

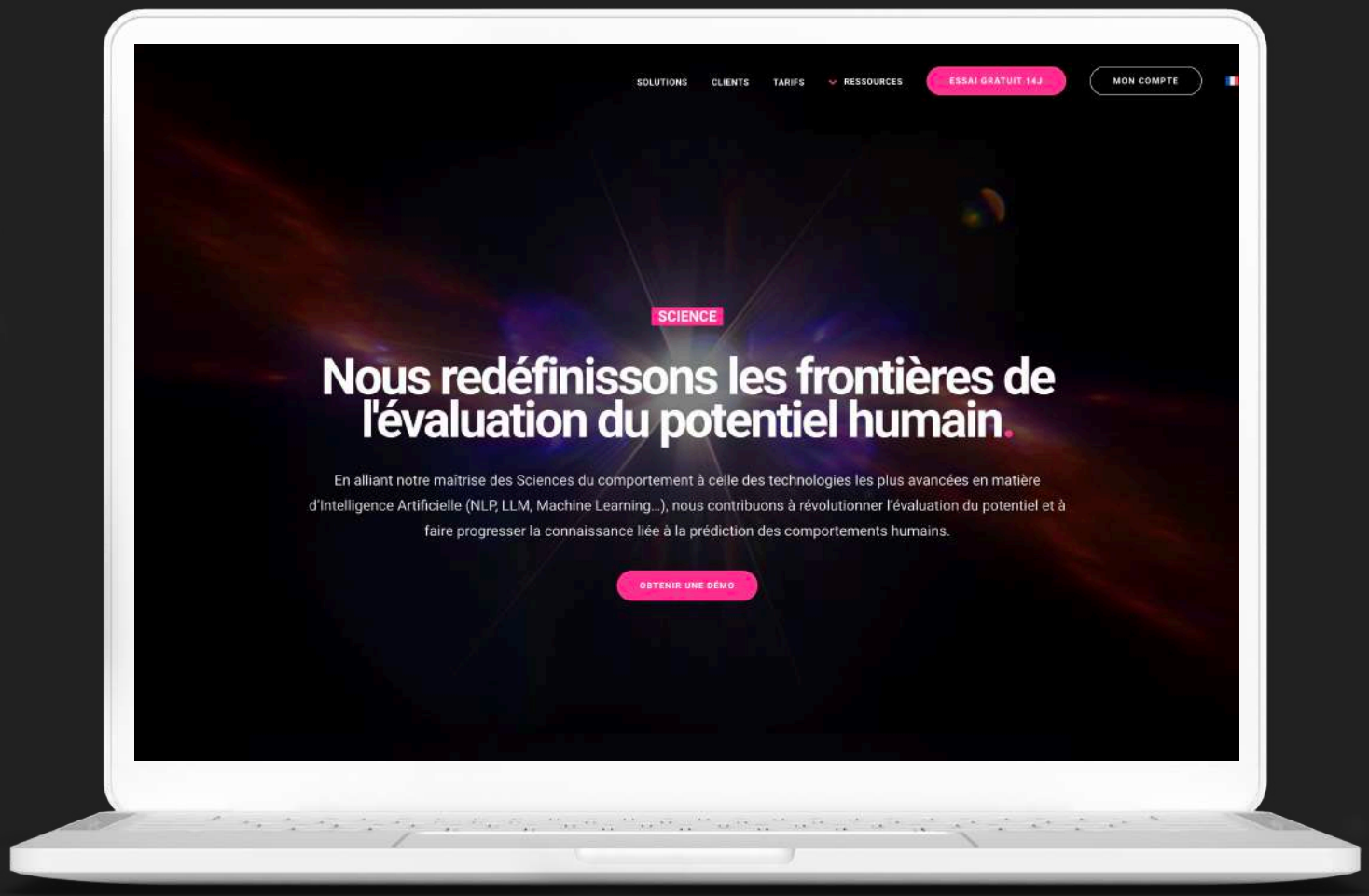
Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806-838. doi: 10.1177/ 0011000006288127

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 a développé une solution de recrutement prédictif permettant aux entreprises de prédire dans quelle mesure les candidats et collaborateurs réussiront et prospéreront dans leur travail. La solution AssessFirst analyse les données de plus de 5 000 000 de profils, qu'ils soient candidats, salariés ou professionnels du recrutement. Plus de 3 500 entreprises utilisent la solution pour augmenter leurs performances jusqu'à 40 %, diminuer leurs coûts de recrutement de 20 % et réduire le taux de turnover de leurs employés de 50 %.