TECHNICAL MANUAL.

Presentation of the development process used to construct AssessFirst assessments.

The Assessment Of The Human Potential By ASSESSFIRST

How we built an integrated model based on the assessment of the personality (SHAPE), the motives (DRIVE) and the aptitudes (BRAIN).

By the AssessFirst's Science & Innovation Team

Index.

Introduction.

This technical manual provides information regarding the design, development and validation of the AssessFirst's questionnaires. It also gives some guidance on the applications and recommended use of the AssessFirst's questionnaires.

AssessFirst's questionnaires have been developed to highlight the full potential of any individual, based on the assessment of his personality (with SHAPE questionnaire), his motivations (with DRIVE questionnaire), and his reasoning skills (with BRAIN questionnaire).

SHAPE questionnaire has 90 questions measuring 20 dimensions of personality. It takes about 11 minutes to complete.

DRIVE questionnaire has 90 questions measuring 20 dimensions of motivations. It takes about 10 minutes to complete.

BRAIN questionnaire has 38 questions measuring 4 reasoning skills, and a general factor. It takes about 18 minutes to complete.

AssessFirst's questionnaires were designed for the age of 18 and over, for use in a work context : preselection, selection, development, career planning…

Item Response Theory (IRT) methodology is used to calculate candidates scores on each cognitive, motivational and personality dimension. We also have a continuous improvement process, so every language version of a questionnaire is updated each year, to maximise its psychometric property.

AssessFirst assessment platform is specifically designed to predict the future success of an individual, thanks to an efficient job matching system based on this principles :
-We assess the individual characteristics which are closely related to work performance.
-We develop smart algorithms to calculate the matches and efficiently generate recommendations.
-We value the candidate experience, because it improves their commitment into the process.

-We make the interpretation accessible, and easily understandable, so the decision-making based on the results is effective.

For further information and questions about the AssessFirst's questionnaires, please refer to the test publisher.

Construction.

This part describes the development methodology used to construct the questionnaires.

DEVELOPMENT BACKGROUND

SHAPE development background

The inventory of the SHAPE questionnaire was developed in 2003 by an occupational psychologist, David Bernard, and his team of psychometric psychologists with the purpose of evaluating someone's personality in a professional context. Many other personality inventories already existed with this goal. But - for the large majority - they were designed in a complex way and addressed to experts. However, more and more human resources professional are looking for tools to make better selection and orientation decisions.

Validity is (or at least should be) the main goal of test developer. This being said, the validity is overestimated if the test users can't size all the subtlety of the scores, the shade of the interpretation, and the inter-dimensions' influence. It's like selling a professional music instrument to a beginner and expect him to create great sounds. To address this limit, test publishers deliver training and guidelines.

On the other side, most easy-to-use questionnaires are based on typological theory of personality, which is simplistic and limit the application to a personal development purpose.

SHAPE was developed so the human resources professionals could have a tool they can use appropriately to make better decisions. It still involves training and other resources, but they have more accessible information to analyse people's personality and to anticipate their job fit.

The second motivation to develop SHAPE was the expansion of the internet. So far, only some questionnaires were online. The idea was not just to replace the paper-pencil form by the online form, but also to use the intelligence of it to improve the assessment. For example, by using all the data collected to create a benchmark, or by adding « bonus » questionnaires to nurture the research and development.

Taking as a starting point the Big Five model - the highly appreciated and scientifically validated psychological concept - the team developed an original model of personality traits called SHAPE that measures 20 facets of the personality. The model focuses principally on the relational, intellectual and emotional human behaviour aspects.

After defining the framework of the model SHAPE, the team of psychologists created a large pool of items in order to capture the total spectrum of human behaviour that was aimed to be measured by the model.

The first version of the questionnaire inventory was constructed in a normative way. It allowed the validation of the structure of SHAPE and to select the most relevant items for each of the 20 personality dimensions.

The development was done by following the guidelines of Gordon (1951). Other important studies taken in strong consideration while designing the questionnaire were the studies of Pytlik Zillig, Hemenover and Dienstbier (2002) that helped to thoroughly cover the emotional, behavioural and cognitive components of each of the Big Five personality traits.

The second version of the questionnaire inventory was thereafter developed in an ipsative form in order to allow the questionnaire to be used in the recruitment context. This reshaped version also helped to significantly neutralise the candidates' inclination to position their responses consciously (Bowen, Martin and Hunt, 2002), the tendency known to psychologists as the bias of social desirability.

The SHAPE items have been continuously improved in order to increase the overall quality of the questionnaire. The items are evaluated scientifically by research psychologists to measure how well they differentiate between individuals on the dimensions.

For example, when detecting that some items in the questionnaire do not measure the particular trait as expected, alternative proposals are developed in order to replace these. The new proposals are first added as “bonus" items in the end of the questionnaire. Those extra items are not taken in account right away in the test scores, but they provide valuable information to the research psychologists about their discrimination properties.

After some trial tests, the results are studied and if the new items are acting better than their original corresponding items in the test, then the old items are replaced with the new ones.

BRAIN development background

BRAIN was developed by AssessFirst in 2020. It was designed to assess the general cognitive ability of an individual—also known as the “g factor”—which is the best single predictor of the capacity to succeed in a professional context, for almost all positions and roles. It was engineered based on 3 design principles, aimed at satisfying the needs and expectations of both clients and candidates:

- A mobile-friendly interface
- An adaptive framework that adjusts to the taker's ability
- A stimulating and engaging experience

BRAIN is a general test, without subdivisions or themes. It requires users to complete a series of logical figures, within a limited amount of time. Instead of a multiple-choice approach, the user builds each answer by combining the available elements on the screen. The test's adaptive structure (Computer Adaptive Testing, or CAT) ensures the level of difficulty of each figure is adapted to the user's real performance, which means each user will benefit from a personalized experience. In brief, the CAT model works as follows:

- The first test item for all users is of medium difficulty
- The following items will be selected based on the user's responses
- If the user answers correctly, the following item will be more difficult
- If the user answers incorrectly, the following item will be less difficult
- The test ends when the user responds consistently to questions of a similar level of difficulty.

DRIVE development background

Appearing in the last century, after those relating to intellectual abilities and personality, the evaluation of motivations is now a major topic in work psychology: what drives a person to engage in an action? The first renowned researchers on motivation competed to develop a universal model of motivation. This was the case with Maslow (2004), McClellan (1961) as well as Herzberg (1968).

Further research has shown that motivation cannot be explained by a single factor, but is in fact multi- facetted. Latham and Pinder (2005) developed an excellent overview of works on the subject, arriving at the following conclusions:

- Satisfaction at work is closely linked to performance (Judge et al, 2001)
- Motivation at work reduces the desire to quit a job (Williams,

Konrad, Scheckler, Pathman, Linzer, McMurray, Gerrity and Schwartz, 2001)
• Motivation at work reduces absenteeism and increases commitment to a position (Wegge, Schmidt, Parkes and van Dick, 2007)

AssessFirst's intention in developing DRIVE was to take advantage of the knowledge revealed by this research, and to put it in service to businesses and individuals with the goal of helping them to better understand individual motivation, and to use it to make better decisions. DRIVE was designed with the objective of identifying the determinants of the satisfaction and commitment of a person in a professional context.

DRIVE was developed by the Science & Innovation team at AssessFirst starting in 2014, and was commercialised at the start of 2016. It is the result of a significant review of literature, studies of existing evaluation tools for motivation, and research into the distinctive factors of committed people who are successful in their role.

Developed after the SHAPE and BRAIN tests, DRIVE has inherited AssessFirst's know-how in terms of the relevance of the evaluated criteria, the simplicity of taking the test and exploiting the results, as well as its psychometric power and algorithmic technology.

THEORETICAL BASIS

Theories behind SHAPE

Theoretical foundations of SHAPE construct are based on the factorial approach.

Factor analysis is a statistical technique which allows the reduction of a large number of correlated variables to a smaller number of “main variables".

It also aims to bring out within the number of observed variables the smaller number of underlying (latent) variables. It does this by attempting to account for the pattern of correlations between the variables in terms of a much smaller number of latent factors. A latent factor is the one that cannot be measured directly, but is assumed to be related to a number of measurable and observable variables.

The psychological theory behind the conception of the SHAPE assessment is the Big Five personality traits model (for detailed information about the development of Big Five model, please refer to Digman, 1990).

The Big Five model is also called the Five Factors model, as the dimensions were derived from factor analyses of a large number of self- and peer reports on personality-relevant adjectives and questionnaire items.

The Big Five identifies five distinct factors as central to personality: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each of these five personality traits describes the frequency or intensity of a person's feelings, thoughts or behaviours compared to other people. It is commonly accepted that everyone possesses all five of these traits to a greater or lesser degree, allowing us to make that comparison between individuals.

We came from the five factor model and then we select the personality scales who were the most related to the work context expectations, after a review of the current personality questionnaires available on the market.

Theories behind DRIVE

The study of motivation has been a critical focus in the last fifty years of research in work psychology. We have been able to draw inspiration from theoretical reflections and field studies conducted on the subject. Some particular works have caught our attention by aligning closely with our primary objective: to identify the main factors of satisfaction and commitment at work. Below is a short summary.

Theory of self-determination - Deci and Ryan

Deci and Ryan's theory of self-determination is one of the key inspirations for the DRIVE questionnaire. The work of these researchers is now considered authoritative in work psychology. Their theory postulates people cannot be motivated by extrinsic incentives over the long term – they must find intrinsic sources of motivation to maintain their commitment to a task.

This theory refers in particular to the fulfilment of 3 basic needs: the need for competence (being capable of doing what is asked), the need for autonomy (being able to independently overcome any difficulties faced, using our creativity) and the need for belonging (being surrounded by people who can be counted on and who will provide support).

This theory can be found in DRIVE in two different ways:
• Of the 20 dimensions that make up the test, 18 are dimensions that refer back to intrinsic
motivation. We have kept 2 extrinsic dimensions because the meta-analysis of Cerasoli, Nickin & Ford (2014) highlighted

that these can also have a positive impact in certain workplace situations.
• DRIVE integrates the 3 fundamental needs expressed in the theory of self-determination.
However, while this model is well-suited to auditing the current level of a person's motivation, it's not necessarily suited to the aim of exploring the different factors for a person to thrive.

Motives Values Preferences Inventory (MVPI) - Joyce and Robert Hogan

This inventory, developed by Joyce and Robert Hogan, was an important benchmark in the development of DRIVE, insofar as these two doctors of psychology investigated and synthesized 80 years of research on the subject. Almost 20 years later, their work still has an excellent reputation because their model is transverse to the theories and experimental research conducted in the field, and it is particularly well-suited to a professional or workplace context (it was designed with this in mind).

Therefore, he 10 MVPI dimensions can be found indirectly in DRIVE.
Multidimensional Theory of Person-Environment Fit
Our research on the work carried out over the last few years led us to investigate deeper into ways of understanding the notion of motivation, particularly through the modelling of different levels of "fit" between a person and their working environment.

The bulk of the research on satisfaction and commitment at work reports the impact of the links between a person and their working environment. The research of Edwards and Billsberry (2010) indicate the different levels of “fit" linked to the satisfaction and commitment of people in professional contexts. They particularly rely on the model created by Jansen and Kristof-Brown (2006) which identifies

5 types of fit:
• The “person-vocation fit"
• The “person-organisation fit"
• The “person-group fit"
• The “person-job fit"
• The “person-person fit"

These levels of fit influence three major indicators: commitment, desire to quit a position, and satisfaction at work. We developed DRIVE with a clear view of promoting these three indicators. This also required us to assimilate existing models to achieve our objective.

Theories behind BRAIN

BRAIN is built on models of the general intelligence factor, or “g factor," such as those developed by Spearman (1904) and Cattel, Horn and Carroll (1997), which are the authority in the field of occupational psychology today.

Unlike its predecessor, which distinguished between four different types of reasoning (verbal, analogical, abstract and numerical), the new BRAIN focuses its analysis on a single general intelligence factor. This evolution is based on two main principles. On one hand, the assessment of specific abilities provides no increase in predictive validity. As demonstrated by Ree, Earles and Teachout (1994), an assessment of g alone is sufficient in predicting job performance, and testing for specific abilities (such as numerical or verbal) does not provide additional information on a candidate's potential for success. Additionally, as demonstrated by Carroll's hierarchical model of intelligence (1993), specific reasoning abilities are strongly intercorrelated, and correlated to a general intelligence factor.

On the other hand, although we cannot deny the value of gathering all possible data on a test-taker's capacity to process information of different types, the excessive duration and arduousness required for such an assessment cannot be justified in light of its predictive power. It is useless to measure additional dimensions, however interesting, as they only contribute to lengthen the test needlessly. We have therefore chosen to prioritize better user experience and shorter test duration.

Validity.

To make better decisions about people on the basis of an assessment, it is important to provide evidence of its validity. It is fundamental and central to effective test application. The aim of validity is to help us to confirm that the assessment in use is really measuring what it is intended to measure and how accurate the results are that we can obtain from it. Furthermore, an assessment is valid for measuring an attribute if variation in the attribute causally produces variation in the measured outcomes.

DEFINITION OF VALIDITY

There are different types of validity evidence and they depend on the characteristics and the purpose of the assessment. AssessFirst's questionnaires have been designed to maximise validity in predicting overall effectiveness at work and key workplace behaviours.

The validity studies on our questionnaires are divided into the following validity subsections:
- Content related validity, which is used to analyse whether the experiment provides adequate coverage of the subject being studied. Here we are focusing on the content validity evidence studies.
- Construct related validity, it refers whether the test actually measures what it is intended to measure (i.e. the construct), and not other variables. Here we are focusing on the construct, structural and convergent validity evidence studies.
- Criterion related validity, which is used to predict future or current performance, correlating test results with another criterion of interest. We are measuring it with the predictive validity evidence studies.
- Cross-cultural equivalence, that helps us to analyse if an assessment is valid to be used on an international basis.

In the following paragraphs we are focusing on each of these validity types, by first describing the concept and then bringing out some of our studies showing validity evidence for our questionnaires.

CONTENT VALIDITY

Content validity, also known as logical validity, is related to the appropriateness of the content of the assessment. It aims to

verify to what extent the test represents all facets of a given construct.
Content validity corresponds also to the quality of the assessment items' formulation, as the whole range of items should be representative for the whole theoretical model.

A content validation study does not rely upon a statistical analysis, but it uses a rationale approach to linking the to-be-measured-construct content to assessment content. The most common way to measure content validity is to ask the experts/ judges of the domain to evaluate the relevance of each item for the pre-linked dimension.

It is also possible to present the range of items to the experts/ judges and ask them to make the link by themselves for each item with the dimension that seems the most suitable. This way of proceeding allows the validation process to be more objective as there is no preliminary connections made and the outcome is more informative that just a declaration of suitability. This latter process of evaluation was prioritised by the research psychologists team of AssessFirst to measure the content validity of SHAPE personality questionnaire.

Study on SHAPE content validity

The study on SHAPE content validity was conducted in 2003.

Twelve judges were selected to participate in the study.
Each judge received a first document (Document A) that detailed the 180 items of the SHAPE inventory and a second document (Document B) presenting the definitions of the 20 dimensions evaluated by the inventory.

The instruction that was presented to the judges was as follows: "The SHAPE inventory has 20 facets. Each facet contains 9 items. For each of the items presented in Document A, you must select in Document B the dimension to which it refers the most. An item can only belong to a single facet."

Table below shows for each facet the number of items over the total (9) that have very high (90 - 100%), high (80 - 90%), correct (70 - 80%) and insufficient (0 - 70%) affiliation with the dimension.

Facets	91 - 100%	81 - 90%	71 - 80%	0 - 70%	Total
Is assertive with others	9	0	0	0	9
Tries to convince others	6	2	1	0	9
Spontaneously approaches others	6	3	0	0	9
Demonstrates diplomacy	8	1	0	0	9
Connects emotionally	7	1	1	0	9
Is open to other people's ideas	9	0	0	0	9
Accepts criticism	9	0	0	0	9
Consults others before making decisions	5	2	1	1	9
Prefers varied tasks	9	0	0	0	9
Is interested in abstract ideas	6	2	1	0	9
Demonstrates inventiveness	9	0	0	0	9
Adapts to change	4	4	1	0	9
Organizes work methodically	9	0	0	0	9
Pays attention to details	4	3	2	0	9
Perseveres when confronted with obstacles	6	3	0	0	9
Goes beyond the assigned tasks	6	1	1	1	9
Is relaxed	5	2	2	0	9
Focuses on the positive	9	0	0	0	9
Controls own feelings	8	1	0	0	9
Seeks stability	5	3	1	0	9
Mean of degree of agreement	**6.95**	**1.4**	**0.5**	**0.1**	**9**

The study shows that for all the facets, the mean of the degree of agreement is very high for 6.95 facets, high for 1.4 facets, correct for 0.55 facets and insufficient for only 0.1 facets.

We note as well that the two items with insufficient degree of agreement were attributed to different facets by the different judges who misidentified them.

Furthermore, for one judge the attribution errors consist in 90% of the cases in the assignment of an item to a facet close from a conceptual point of view (i.e. connected to the same factor).

Study on DRIVE content validity

The study on DRIVE content validity was conducted in 2015.

Ten judges were selected to participate in the study. Each judge received a first document (Document A) that detailed the 180 items of the DRIVE inventory and a second document (Document B) presented the definitions of the 20 dimensions evaluated by the inventory.

The instruction that was presented to the judges was as follows: "The DRIVE inventory has 20 facets. Each facet contains 9 items. For each of the items presented in Document A, you must select in Document B the dimension to which it refers the most. An item can only belong to a single facet."

Table below shows for each facet the number of items over the total (9) that are very high (> 90%), high (80 - 90%), correct (70 - 80%) and insufficient (< 70%).

Facets	91 - 100%	81 - 90%	71 - 80%	0 - 70%	Total
Create new things	9	0	0	0	9
Excel everyday	8	1	0	0	9
Worry about aesthetics	7	1	1	0	9
Analyse data	7	1	1	0	9
Meet new people	7	1	1	0	9
Have clearly defined tasks	7	2	0	0	9
Worry about quality	6	2	1	0	9
Having influence	8	0	1	0	9
Having autonomy	9	0	0	0	9
Working in a team	8	1	0	0	9
Having a positive impact on the world	7	2	0	0	9
Working in a fun environment	7	0	2	0	9
Developing in a reassuring environment	7	1	1	0	9
Working in a disciplined manner	8	0	1	0	9
Maintaining personal balance	8	1	0	0	9
Receiving compensation	6	2	1	0	9
Having attractive remuneration	5	2	2	0	9
Achieving success regularly	8	1	0	0	9
Helping others	6	2	1	0	9
Being recognized by others	7	1	1	0	9
Mean of degree of agreement	**7.25**	**1.05**	**.70**	**0**	**9**

The study shows that over all the 20 facets, the mean of the degree of agreement is very high for 7.25 items per facet, high for 1.05 items, correct for 0.7 facets. There are no insufficient items found in the facets.

CONSTRUCT RELATED VALIDITY

Construct validity allows us to find out whether the test actually measures the intended theoretical construct/trait or, partly or mainly, something else. This type of validity overlaps with some of the other aspects of validity, as all validity evidence contributes to an understanding of an assessment instrument's construct validity.

Construct validity is important because it impacts how an assessment score is interpreted. If a questionnaire claims to measure a specific personality trait, how do we know it actually measures this trait and not anything else? If a questionnaire is supposed to measure a specific trait but in reality it does not, then any interpretation of that score would be not correct and may cause damage to the test taker as well as for the organization.

Construct validity does not concern a simple, factual question of whether an assessment measures a trait. Instead it contains complex investigations of whether test score interpretations are consistent with a nomological network involving theoretical and observational terms (Cronbach & Meehl, 1955).

There is no single method of determining the construct validity of a test, but different methods and approaches are combined to present the overall construct validity of an assessment.

To measure the construct validity of SHAPE and DRIVE questionnaires, the AssessFirst's Research and Development team has employed the following methods:
- Item-dimension saturation,
- Inter-dimension correlation.

Study on SHAPE construct validity

The latest studies on SHAPE construct validity were conducted in 2016.

Item-dimension saturation

A study on item-dimension saturation proceeded using Pearson's correlation (r). There were 43 217 test scores analysed.

Table below shows the results of this analysis, where the saturation of the items varied from .23 ≤ r ≤ .74, between dimensions and their respective items.

SHAPE item-dimension saturation (N=43 217).

Facets	I1	I2	I3	I4	I5	I6	I7	I8	I9
Is assertive with others	.39	.31	.61	.45	.55	.40	.67	.46	.27
Tries to convince others	.51	.34	.36	.58	.23	.47	.30	.45	.63
Spontaneously approach others	.36	.60	.55	.37	.65	.35	.63	.40	.39
Demonstrates diplomacy	.36	.40	.29	.27	.41	.42	.78	.43	.34
Connects emotionally	.42	.38	.34	.48	.59	.38	.50	.41	.33
Is open to other people's ideas	.37	.29	.30	.29	.28	.48	.59	.42	.31
Accepts criticism	.36	.46	.44	.52	.39	.66	.47	.57	.28
Consults others before mak..	.30	.26	.37	.42	.40	.52	.57	.47	.49
Prefers varied tasks	.38	.46	.50	.57	.57	.32	.34	.35	.25
Is interested in abstract ideas	.43	.42	.70	.36	.41	.36	.27	.36	.47
Demonstrates inventiveness	.53	.45	.33	.57	.32	.54	.36	.33	.43
Adapts to change	.44	.74	.47	.42	.38	.34	.27	.67	.33
Organizes work methodically	.39	.44	.39	.60	.45	.33	.38	.44	.48
Pays attention to details	.61	.67	.32	.36	.36	.29	.44	.57	.54
Perseveres when confronted	.31	.42	.36	.40	.23	.50	.65	.29	.44
Goes beyond the assigned	.66	.34	.37	.48	.37	.30	.38	.52	.31
Is relaxed	.43	.43	.30	.56	.34	.47	.49	.36	.46
Focuses on the positive	.51	.39	.48	.42	.44	.32	.44	.40	.57
Controls own feelings	.56	.31	.31	.30	.40	.49	.44	.40	.60
Seeks stability	.52	.46	.40	.57	.58	.34	.45	.34	.36

Inter-dimension correlation

A study on SHAPE questionnaire inter-dimension correlations (coefficient Pearson r) was conducted on 43 217 test scores. The results are shown on table below.

SHAPE inter-dimension saturation (coefficient Pearson r) (N=43 217).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	.40	.15	-.11	-.14	-.06	-.07	.05	.03	.14	.12	.12	-.17	-.37	-.12	.36	-.17	-.04	-.27	-.21
2		1	.01	-.02	-.14	-.03	-.04	.00	-.04	.20	.06	.06	-.21	-.19	-.06	.23	-.14	-.22	-.32	-.15
3			1	.17	.10	-.10	.06	-.05	-.20	-.13	-.15	.11	-.18	-.15	-.14	.10	-.12	.11	-.31	-.16
4				1	.16	.09	.09	.06	-.07	-.14	-.19	-.10	-.10	-.10	-.10	-.36	-.06	.09	.04	.00
5					1	.25	.27	.22	-.07	-.36	-.14	-.03	-.12	-.12	-.16	-.17	-.23	.14	-.10	-.03
6						1	.29	.26	-.11	-.05	-.04	.02	-.18	-.14	-.43	-.15	-.14	-.11	-.03	.00
7							1	.36	-.07	-.23	-.14	.02	-.15	-.21	-.25	-.14	-.29	.07	-.11	-.04
8								1	-.10	-.10	-.10	.00	-.19	-.13	-.38	-.07	-.23	.01	-.19	.02
9									1	.16	.04	.19	-.15	-.12	.00	.12	.00	-.01	-.06	-.15
10										1	.24	.03	.01	-.10	.00	.11	-.01	-.18	-.02	-.21
11											1	.13	-.14	-.14	-.11	.24	-.04	-.19	-.08	-.10
12												1	-.27	-.52	-.08	.30	-.19	.18	-.26	-.45
13													1	.45	.14	-.19	.05	-.26	.17	.25
14														1	.23	-.40	.11	-.14	.23	.25
15															1	.06	.16	-.09	.08	.05
16																1	-.14	.04	-.28	-.34
17																	1	.02	.37	.04
18																		1	-.21	-.12
19																			1	.28
20																				1

The inter-dimension correlations range from -.52 to .45, which allows us to confirm that SHAPE has an acceptable level of consistency.

Over all, the dimensions have a low correlation. Most correlated dimensions are the pairs of “Perseveres when confronted with obstacles" and "Is open to other people's ideas", (r = -.43), “Adapts to change" and “Pays attention to details" (r = -.52), and “Organizes work methodically" and “Pays attention to details" (r = .45).

Study on DRIVE construct validity

The latest studies on the DRIVE construct validity were conducted in 2017.

Item-dimension saturation

A similar study to SHAPE on item-dimension saturation proceeded using Pearson's correlation (r). There were 4 550 test scores analysed in 2016.

The results of this analysis, where the saturation of the items varied from .37≤ r ≤ .83, between dimensions and their respective items, with a mean of .44 and median of .47.

Inter-dimension correlation

A study on DRIVE questionnaire inter-dimension correlations (coefficient Pearson r) was conducted on 4 550 test scores. The results are shown on table below.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	.21	-.10	.04	.07	-.25	-.30	.28	.06	.08	.12	-.12	-.24	-.10	-.33	-.11	-.26	-.01	-.10	-.25
2		1	.05	.10	.02	-.04	.08	.12	-.06	-.15	-.02	-.32	-.03	-.26	-.10	-.15	-.09	.20	-.21	-.33
3			1	.05	.05	.23	.31	-.19	-.07	.04	.07	-.22	.03	-.21	.21	-.37	-.12	-.15	.05	-.28
4				1	-.26	.15	.18	.08	-.19	-.14	-.04	-.30	.07	-.09	.08	-.09	-.04	.08	-.15	-.07
5					1	-.13	-.23	-.19	-.17	.37	.22	.06	-.22	-.03	-.06	-.16	-.29	-.32	.26	-.06
6						1	.36	-.44	-.07	-.08	-.09	-.09	.26	.11	.35	-.22	-.01	-.08	-.13	-.15
7							1	-.28	-.09	-.27	-.25	-.24	.23	-.22	.32	-.12	.08	.15	-.10	-.08
8								1	.05	-.11	-.13	-.05	-.16	-.14	-.36	.13	.01	.21	-.21	.02
9									1	-.13	-.04	.17	-.15	.13	-.13	-.04	.05	-.12	-.03	-.12
10										1	.23	.05	-.21	-.04	-.11	-.23	-.30	-.26	.32	-.02
11											1	.11	-.25	.01	-.03	-.29	-.45	-.40	.28	-.04
12												1	-.12	.28	-.15	.03	-.13	-.33	.12	.25
13													1	.07	.26	.10	.26	.12	-.24	-.07
14														1	.07	-.05	-.01	-.24	.03	.04
15															1	-.10	.05	-.03	.02	-.22
16																1	.43	.29	-.20	.28
17																	1	.36	-.28	.06
18																		1	-.33	-.01
19																			1	.05
20																				1

The inter-dimension correlations range from -.45 to .43, which allows us to confirm that DRIVE has a good level of consistency.

Over all, the dimensions have a low correlation. Most correlated dimensions are the pairs of “Having attractive

remuneration" and "Having a positive impact on the world ", (r = -.45), “Having attractive remuneration" and “Receiving compensation" (r = .43), “Have clearly defined tasks" and “Having Influence" (r = -.44).

CONVERGENT VALIDITY

Convergent validity refers to the degree to which two measures of constructs that theoretically should be related, are indeed related. This type of validity evidence can be established if two similar constructs correspond with one another, showing that the assessment of a concept is highly correlated with other tests designed to measure theoretically similar concepts.

To measure the convergent validity for SHAPE personality questionnaire, we are comparing it with the NEO PI-R assessment. The NEO PI-R is a classic test and an essential one when talking about personality. The model of the questionnaire is solid, well established, and measures what one might call the pure personality. As both NEO PI-R and SHAPE are based on the Big Five personality theory, we can study their correlations in order to evaluate the convergent validity of SHAPE.

Study on SHAPE convergent validity

A study of convergent validity between SHAPE and NEO PI-R was conducted in 2014 on a sample of 258 people. 30% of them were men and 70% were women. The average age was 27 years, and varied from 16 years to 62 years. They were selected on voluntary basis and composed a large scale of social and occupational backgrounds.

Each participant completed the two questionnaires online and received afterwards brief written feedback.

Here are the correlation coefficients of SHAPE and NEO PI-R obtained for Big Five dimensions:

Extraversion	.73
Agreeableness	.66
Openness	.60
Conscientiousness	.65
Neuroticism	.59

We conclude that the relationships between the tests are strong enough to validate a similar base of construct. On the other hand, the differences existing between the two tests are explained by the more professional orientation of SHAPE compared to the NEO PI-R which is more of a general approach of personality evaluation.

In detail, we observe a stronger connection with the extraversion dimension, which is understandable as theoretically the dimension "Extraversion - Introversion" is very important in the work context and has therefore been maximized in the SHAPE questionnaire. Next we find Agreeableness and Conscientiousness which also present prominent behaviours in the work context. Finally comes Openness and Neuroticism which are also necessary to understand the human behaviours at work, however less important and of which only the professional context aspects have been retained.

DIVERGENT VALIDITY

Divergent validity is used to determine if a test is too similar to another test. If a test is found to correlate too strongly with another test then it suggests that the tests are measuring the same thing and are too alike to be considered different. Divergent validity tests whether an assessment that is not supposed to be related are actually unrelated.

To measure the divergent validity for SHAPE personality questionnaire, it is necessary to first establish convergent validity, before testing it for divergent validity. Therefore, we are using the previous study described in the chapter “Study on convergent validity" and analyse the correlations of non corresponding NEO PI-R construct, to show the evidence of discriminant validity.

Study on SHAPE divergent validity

A study of divergent validity between SHAPE and NEO PI-R was conducted in 2014 on a sample of 258 people. 30% of them were men and 70% were women. The average age was 27 years, and varied from 16 years to 62 years. They were selected on voluntary basis and composed a large scale of social and occupational backgrounds.

Each participant completed the two questionnaires online and received afterwards brief written feedback. The correlations between the measured traits of the two tests are shown in the Table X.X.

Correlations between NEO PI-R and SHAPE dimensions (N=258).

NEO PI-R dimensions	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
Anxiety	-.19	-.03	.09	-.03	.25	.10	.30	.31	-.19	.01	-.07	-.22	.16	.20	-.06	-.15	-.49	-.20	-.08	.32
Hostility	.13	.19	.02	-.25	.04	-.05	.13	.05	-.04	-.04	.14	.04	-.01	.05	.12	.11	-.32	-.18	-.24	.11
Depression	-.24	.01	-.05	-.05	.23	.15	.23	.23	-.16	.04	.10	-.11	.03	.14	-.03	-.20	-.31	-.33	.07	.27
Self-consciousness	-.30	-.02	-.14	.04	.27	.22	.20	.33	-.17	.11	.03	-.20	.01	.18	-.07	-.22	-.35	-.23	.10	.26
Impulsiveness	.21	.11	.23	.06	.13	-.07	.05	.09	.08	-.07	.06	.23	-.23	-.24	.06	.21	-.18	.09	-.42	-.36
Vulnerability to Stress	-.26	.00	.05	.02	.28	.25	.31	.31	-.14	-.08	.00	-.13	-.07	.13	-.23	-.16	-.39	-.11	-.01	.27
Warmth	.11	-.11	.43	.27	.35	.09	.00	.06	-.10	-.13	-.20	-.15	-.04	-.03	-.08	-.04	-.14	.22	-.30	-.27
Gregariousness	.15	.04	.42	.13	.18	.11	.16	.29	-.05	-.20	-.28	.06	-.13	-.11	-.19	.00	-.22	.11	-.31	-.14
Assertiveness	.64	.32	.24	-.05	-.14	-.25	-.30	-.25	.13	-.04	.04	.07	-.05	-.23	.08	.28	.09	.10	-.40	-.33
Activity	.41	.07	.22	-.04	-.10	-.32	-.23	-.25	.30	-.07	.05	.26	.00	-.22	.15	.41	-.04	.20	-.46	-.35
Excitement Seeking	.26	.06	.14	-.03	-.11	.03	.05	-.05	.10	-.05	.11	.25	-.21	-.21	-.03	.23	.05	.07	-.22	-.40
Positive Emotion	.22	-.10	.31	.15	.13	-.04	-.02	-.02	.00	-.14	-.06	.13	-.15	-.21	-.05	.19	-.01	.51	-.41	-.41
Openness to Fantasy	.04	-.08	-.04	.00	.05	.11	-.05	-.07	-.09	.20	.46	.11	-.25	-.26	-.11	.11	-.02	.04	-.02	-.13
Openness to Aesthetics	.10	-.02	.10	.13	.08	.15	-.09	.05	-.18	.15	.26	.04	-.27	-.19	.03	.03	-.08	-.02	-.13	-.14
Openness to Feelings	.14	.04	.20	.05	.20	.27	-.02	.07	-.13	.01	.12	-.02	-.17	-.06	-.05	.15	-.23	.05	-.38	-.20
Openness to Actions	.28	-.07	.02	-.05	-.09	.03	-.11	-.10	.14	.04	.30	.34	-.23	-.38	-.03	.25	.16	.09	-.16	-.39
Openness to Ideas	.05	.09	-.06	.05	-.17	.03	-.20	-.13	-.08	.42	.37	.00	-.28	-.16	.07	.12	.09	-.12	.06	-.16
Openness to Values	.00	-.15	-.07	.04	.02	.17	-.04	.05	.01	.11	.06	.18	-.12	-.18	-.13	-.02	.11	.11	.10	-.19
Trust	.02	-.11	.17	.22	.18	.19	-.01	.08	.02	-.07	-.16	.06	-.06	-.09	-.31	-.04	.00	.34	-.15	-.27
Straightforwardness	-.23	-.26	.08	.00	.26	.22	.12	.17	.08	-.01	-.15	-.14	.12	.10	-.27	-.12	-.07	.05	-.01	.07
Altruism	-.04	-.15	.21	.18	.35	.16	-.02	.15	-.06	-.14	-.19	-.22	-.02	.06	-.05	-.02	-.06	.08	-.12	-.13
Compliance	-.39	-.34	-.02	.18	.31	.26	.16	.19	-.09	-.03	-.13	-.07	.11	.01	-.22	-.26	-.03	.12	.16	.14
Modesty	-.29	-.33	-.09	-.01	.34	.25	.15	.20	.02	-.05	-.12	-.14	.09	.12	-.10	-.21	-.15	-.06	.22	.22
Tender-mindedness	-.16	-.20	.14	.12	.43	.37	.14	.30	-.14	-.06	-.11	-.15	-.04	-.03	-.20	-.20	-.20	.07	-.14	.11
Competence	.11	.05	-.07	-.15	-.12	-.14	-.21	-.14	.04	.06	-.11	-.10	.40	.25	.14	.08	.06	-.12	-.06	-.02
Order	.03	-.11	-.08	-.11	-.11	-.17	-.10	-.03	-.06	-.01	-.13	-.14	.60	.41	.14	-.07	-.03	-.19	.00	.06
Dutifulness	.04	-.02	-.04	-.05	-.11	-.20	-.16	-.06	-.02	.02	-.22	-.16	.31	.30	.29	.08	.08	-.06	.01	-.08
Achievement Striving	.25	.08	-.03	-.17	-.10	-.21	-.27	-.21	.10	.03	-.01	.04	.25	.11	.34	.30	-.05	-.09	-.25	-.14
Self-Discipline	.13	-.03	-.10	-.15	-.09	-.28	-.18	-.17	.04	-.01	-.14	-.12	.47	.28	.31	.17	.00	-.04	-.08	-.08

The table shows us that most of the NEO PI-R dimensions are moderately correlated (> .30 or < .-30) with one or many SHAPE dimensions (D1-D20). However, there are three NEO PI-R dimensions that are not correlated significantly with any of SHAPE dimension. These are “Openness to aesthetics" (-.27 < r > .26), “Openness to values" (-.19 < r > .17) and “Straightforwardness" (-.27 < r > .26).

CRITERION RELATED VALIDITY

Predictive validity

Predictive validity evidence is particularly applicable when one wishes to make an inference from an assessment score about the

test taker's position on another independently evaluated criterion variable at a later date.

In the predictive validity evidence studies, we investigate how much an assessment that is designed to forecast the individual's potential correlates with his future work position performance (for example with the supervisor performance ratings). Such an assessment would have predictive validity if the observed correlation were statistically significant.

To study the predictive validity of AssessFirst's questionnaires, we investigate how much they forecasted the potential to perform in a particular position, in a particular company.

Our predictive validity evidence studies can be divided into the following three steps: preparation, execution, verification.

To find more details about our methodology and results, please have a look on the document « predictive validity studies ».

CROSS-CULTURAL EQUIVALENCE

Cross-cultural equivalence is a type of validity that provides evidence that an assessment is valid to be used on an international basis. It refers to the measurement level at which scores can be compared across cultures. In other words, the validity shows whether the test scores obtained in different cultural populations can be interpreted in the same way across these populations. It is a crucial point when an instrument has been translated or adapted from a non-local context. Without this it is not possible to generalise findings from one country or language version to another.

As the SHAPE personality questionnaire English version is an adaptation from the original French version and is indeed meant to be used internationally, it is very important to study its cross-cultural equivalence.

To ensure the cross-cultural equivalence of SHAPE questionnaire, we are following the classification proposed by Van de Vijver and Poortinga (2005):

Structural equivalence:

Structural equivalence implies the existence of similarity of data structures across cultures. This equivalence assesses if the set of observed indicators (e.g. questionnaire items) has the same pattern or structure of existing and non-existing relationships (e.g. factor loadings) with the construct to be measured across cultures. It is not fully necessary that these relationships have exactly the same

strength but that the same set of questions is related to same concepts in each culture.

Measurement unit equivalence:

While structural equivalence does not indicate that respondents from different cultures assign the same meaning to questions, the measurement unit equivalence includes the structural equivalence and additionally assumes that the relationship between observed indicators and latent concepts is equal across different groups.
In other words, measurement equivalence implies the equality of the measurement units on which a trait is assessed across cultural groups.
This level of equivalence implies that the instrument measures the same latent construct in all of the cultural groups under investigation. Thus, measurement equivalence represents a necessary condition for comparison of difference scores (e.g. mean-corrected scores) across countries. It also enables valid comparison of relationships of the latent variable with other variables of interest (Steenkamp & Baumgartner, 1998).

Scalar equivalence:

In order to establish complete cross-cultural equivalence and to enable full comparison of country scores, it is necessary that the scales of the latent construct have the same origin.
When measures also have a common origin across groups, they are considered to have scalar equivalence.
This level of equivalence can be obtained when two assessments have the same measurement unit and the same origin (i.e. raw scores have the same meaning and can be compared across groups).

Study on SHAPE French and English cross-cultural equivalence

Structural equivalence evidence:

A study on structural equivalence between SHAPE original version in French and adapted version in English was done in April 2017.

The item-dimension correlations were compared by their means between the two tests to analyse whether the test items are following the same pattern of variance.

The results in Table below show excellent invariance in the two test construct structures (mean difference .03). This study confirms the structural equivalence evidence between SHAPE French and English versions.

Table : SHAPE (English vs French) mean item-dimension saturation comparisons.

Facets	Mean (EN)	Mean (FR)	Difference
Is assertive with others	.51	.46	.05
Tries to convince others	.49	.43	.06
Spontaneously approaches others	.49	.48	.01
Demonstrates diplomacy	.43	.41	.02
Connects emotionally	.47	.43	.04
Is open to other people's ideas	.44	.37	.07
Accepts criticism	.46	.46	.00
Consults others before making decisions	.47	.42	.05
Prefers varied tasks	.48	.42	.06
Is interested in abstract ideas	.44	.42	.02
Demonstrates inventiveness	.47	.43	.04
Adapts to change	.43	.45	.02
Organizes work methodically	.46	.43	.03
Pays attention to details	.48	.46	.02
Perseveres when confronted with obstacles	.44	.40	.04
Goes beyond the assigned tasks	.47	.41	.06
Is relaxed	.46	.43	.03
Focuses on the positive	.45	.44	.01
Controls own feelings	.44	.42	.02
Seeks stability	.39	.45	.06
Mean	**.46**	**.43**	**.03**

Measurement unit equivalence evidence:

To provide the evidence of the measurement unit equivalence between the original version in French and the adapted English versions, we must analyse whether the questionnaire items are understood in a similar way in different languages.

One possible way to measure the measurement unit equivalence is to compare the content validity of both assessments. As seen above, the content validation study uses a rationale approach to linking the to-be-measured-construct content to assessment content.

The comparison between SHAPE French and English content validity studies (Tables 4.1. and 4.2 above) shows us that over all the 20 facets, the mean of the degree of agreement for English version is “very high” on 6.55 items per facet, which meets the results for French version (6.95 items per facet). SHAPE English facets are

attributed “highly” on 1.6 items, similar to SHAPE French on 1.4 items. “Correct” attribution is on 0.85 items per facet for English version, and 0.55 items for French version. There are no “insufficient” items found in the facets for English version, while for French it is only 0.1 items.

Comparison of these two content validity studies allows us to confirm the cross-cultural items attribution similarity in the SHAPE English and French versions.

Scalar equivalence evidence:

To demonstrate the scalar equivalence evidence between the SHAPE questionnaire English version and the original French version, an item comparison study was conducted to evaluate the amount of cultural adaptations made during the translation-adaptation process.

The aim of the study was to measure whether the items share the same meaning and can be compared across different language groups.

In this study, SHAPE questionnaires in English and in the original French version were compared item by item. The amount of differences (slight or important) in the item meanings were recorded.

The results showed that 14 out of a total of 90 items were translated so that the meaning of the item had changed during the adaptation process.
Eight of those changes were slight, meaning that the item had gone through some modifications in wording, but the core meaning of the item stayed the same.

Six items contained more important changes, with a different meaning of a phrase being given. However, the items were attributed respectively to the same dimension in both of the versions.

It is important to note that equivalence of the parameters for all items is not necessary for substantive analyses to be meaningful. Cross-cultural comparison can be made in a valid way if at least two items per construct are equivalent (Kankaras & Moore, 2010).

This is the case for SHAPE questionnaire, as each scale is measured within 9 different items and the items that have gone through (slight or important) changes as described above, are distributed over the 20 dimensions without more than two modified items per scale.

Reliability.

To make better decisions about people on the basis of an assessment, it is important to provide evidence of its validity. It is fundamental and

Definition of reliability

Reliability of an assessment is concerned with how precisely the instrument measures particular traits. It provides an index of how exact and error-free an assessment is in measuring the desired constructs. Information on the reliability is an important prerequisite for the test user to draw accurate conclusions from test scores.

The observed scores on the assessment are intended to give an approximation of the person's true scores. The higher the reliability of an assessment, the less is the error and the more likely the observed scores are an accurate reflection of the individual's true scores. Otherwise, if the measured scores are unreliable, they provide less precise and less accurate information on the true scores.

Reliability in IRT is defined as a function that is conditional on the scores of the measured latent construct. Precision of measurement differs across the latent construct continuum and can be generalized to the whole target population. In IRT, measurement precision is often depicted by the information curves. These curves can be treated as a function of the latent factor conditional on the item parameters. They can be calculated for an individual item (item information curve) or for the whole test (test information curve). The test information curve can be used to evaluate the performance of the test (An et Yung, 2014).

To study the reliability of the personality questionnaire SHAPE, we are focusing on the following types of reliabilities:
- Test-retest reliability
- Internal consistency reliability

The studies below were conducted for the SHAPE original version in French and its adapted version in English.

Sources of error affecting reliability

Assessment scores can contain errors of measurement that may affect the reliability. Errors of measurement are liable to influence the result of the investigations.

The different sources of error that typically occur are related to the following:
- Individual state, like the mood, motivation and well-being of the participant while taking the assessment.
- Environmental conditions, like the noise, temperature and presence of others during the assessment.
- Administration mode, including the degree and consistency of standardization.
- Scoring processes.

Maximizing Reliability

The primary development aim of SHAPE questionnaire was to develop a tool with a high validity to predict the professional performance outcomes. In order to achieve it, the tool also needs to be reliable.

To achieve this aim, specific steps were taken to ensure high reliability:
- Negatively phrased and keyed items were avoided as much as possible.
- Questionnaire instructions were standardized for all participants.
- Items were not included if they had low reliability as well as validity.
- Items were selected based on their mean endorsement value from the normative trial to ensure the items were equally attractive to respondents.
- Items were written and reviewed against clear criteria.

Standard error of measurement

When a test user receives a score based on the assessment, the user makes inferences, communicates and takes decisions based that score. Therefore, it is important for a test user to have an understanding of the of error possibilities around the score and know how likely it is to contain the individual's true score. To do this the standard error of measurement (SEM) is calculated.
SEM is closely related to the reliability of the assessment. As a test's reliability goes up, the standard error of measurement goes down and the more precisely the test thus measures the construct.

In classical test theories, the reliability is a one-number summary of test precision, and there is a corresponding single standard error of measurement value that is used for any test

score. In IRT, test precision is conceptualized as something called Information, which is conditional on the trait level being measured.

The IRT based assessments permit the calculation of conditional estimates of measurement precision and generate item and test information curves that more accurately reflect reliability of measurement across all levels of the underlying trait.

Formula of standard error of measurement

3.5. Formula of standard error of measurement
In IRT, the standard error of measurement of a scale is equal to the inverse square root of information at every point along the train continuum and is calculated with the following formula:

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

where:
SE(θ)= standard error of measurement,
I(θ)= test information,
evaluated at given level of the underlying trait θ.

Thus, scales that generate more information yield lower standard errors of measurement.
However, the error is not a static characteristic of an IRT based test. Reliability varies across ability level, and depends specifically on how well the items match the subjects (i.e., on information).

TEST-RETEST RELIABILITY

Test-retest reliability describes the stability of test scores over time. It measures the degree to which test scores are consistent from one test administration to the next.

To assess this reliability, the Pearson Product-Moment correlation coefficient (r) is mainly used to measure the relationship between the group scores on a scale on one occasion and a second occasion.

The test-retest reliability can be affected by the length of time between administrations of the tests. If the interval is too short (i.e. two days), the similarity of scores may simply be explained by individuals' ability to recall their recent responses. If the interval is too long (e.g. 12 months), the dissimilarity in the

scores may be due to actual changes that have occurred in the individuals over time (e.g. birth of a child, change in career).

Although there is not a hard-and-fast rule, experts recommend that the interval between the two occasions of personality testing be at least one week but no more than three months (Carducci, 2009).

Test-retest reliabilities of .70 and above are acceptable levels of reliability.

Study on SHAPE test-retest reliability

The study on test-retest reliability was conducted in April 2017.

Table below provides the test-retest reliability of SHAPE questionnaire administered at a 3 month interval.

The 20 dimensions of SHAPE questionnaire demonstrate excellent test-retest reliabilities with coefficients ranging from .74 to .87 and a median reliability coefficient of .80.

Test-retest reliability of SHAPE (N=1,170).

Facets	(Pearson R)
Is assertive with others	.86
Tries to convince others	.80
Spontaneously approaches others	.87
Demonstrates diplomacy	.81
Connects emotionally	.78
Is open to other people's ideas	.77
Accepts criticism	.77
Consults others before making decisions	.74
Prefers varied tasks	.79
Is interested in abstract ideas	.79
Demonstrates inventiveness	.82
Adapts to change	.76
Organizes work methodically	.82
Pays attention to details	.84
Perseveres when confronted with obstacles	.79
Goes beyond the assigned tasks	.79
Is relaxed	.79
Focuses on the positive	.84
Controls own feelings	.86
Seeks stability	.80
Median	**.80**
Min	**.74**
Max	**.87**

Study on DRIVE test-retest reliability

The study on test-retest reliability was conducted in April 2017.

Table below provides the test-retest reliability of DRIVE questionnaire administered at a 3 month interval.

The 20 dimensions of DRIVE questionnaire demonstrate good test-retest reliabilities with coefficients ranging from .67 to .83 and a median reliability coefficient of .74.

Test-retest reliability of DRIVE (N=537).

Facets	(Pearson R)
Create new things	.76
Excel everyday	.72
Worry about aesthetics	.69
Analyse data	.73
Meet new people	.75
Have clearly defined tasks	.79
Worry about quality	.74
Having influence	.77
Having autonomy	.67
Working in a team	.74
Having a positive impact on the world	.71
Working in a fun environment	.77
Developing in a reassuring environment	.70
Working in a disciplined manner	.69
Maintaining personal balance	.83
Receiving compensation	.74
Having attractive remuneration	.71
Achieving success regularly	.75
Helping others	.77
Being recognized by others	.71
Median	**.74**
Min	**.67**
Max	**.83**

Study on BRAIN test-retest reliability

The study of the test-retest reliability of the new BRAIN questionnaire (released in 09/06/2020) will be conducted at the end of 2021. We are gathering data.

The previous version of the BRAIN questionnaire demonstrated very good test-retest reliabilities with coefficients ranging from .77 to .89. in a study conducted in April 2017 on 511 people after a 12 months interval.

INTERNAL CONSISTENCY

Internal consistency measures consistency of results across all items within an assessment. It assesses whether several items that propose to measure the same general construct produce similar scores. Internal consistency is usually measured with Cronbach's alpha.

However, it is noteworthy that it is difficult to obtain high internal consistencies on the questionnaires with forced-choice responses, as this format distorts the internal consistency of instruments (Brown and Maydeu-Olivares, 2013).

Therefore, as the current version of SHAPE and DRIVE questionnaires is ipsative, the internal consistency is biased as a method for estimating its reliability.

The following studies are presenting the internal consistency for SHAPE and DRIVE questionnaires in normative versions.

Study on SHAPE internal consistency

The study on internal consistency reliability was conducted in 2004.

The table below provides the internal consistency (Cronbach’s Alpha) of the 20 dimensions of SHAPE personality traits.

Internal consistency reliability of SHAPE normative version (N=982).

Facets	(Cronbach's alpha)
Is assertive with others	.79
Tries to convince others	.82
Spontaneously approaches others	.76
Demonstrates diplomacy	.75
Connects emotionally	.68
Is open to other people's ideas	.70
Accepts criticism	.69
Consults others before making decisions	.82
Prefers varied tasks	.83
Is interested in abstract ideas	.89
Demonstrates inventiveness	.91
Adapts to change	.75
Organizes work methodically	.73
Pays attention to details	.80
Perseveres when confronted with obstacles	.81
Goes beyond the assigned tasks	.77
Is relaxed	.93
Focuses on the positive	.82
Controls own feelings	.76
Seeks stability	.76
Median	**.79**
Min	**.68**
Max	**.93**

With the exception of two dimensions, Connects emotionally (.68) and Accepts criticism (.69), the internal consistency coefficients were all above .70, showing adequate reliability results of SHAPE questionnaire.

Study on DRIVE internal consistency

The study on internal consistency reliability was conducted in 2015.

The table below provides the internal consistency (Cronbach's Alpha) of the 20 dimensions of DRIVE motivation factors.

Internal consistency reliability of DRIVE normative version (N=332).

Facets	(Cronbach's alpha)
Create new things	.77
Excel everyday	.71
Worry about aesthetics	.68
Analyse data	.70
Meet new people	.75
Have clearly defined tasks	.87
Worry about quality	.77
Having influence	.74
Having autonomy	.69
Working in a team	.79
Having a positive impact on the world	.72
Working in a fun environment	.72
Developing in a reassuring environment	.72
Working in a disciplined manner	.77
Maintaining personal balance	.81
Receiving compensation	.71
Having attractive remuneration	.75
Achieving success regularly	.82
Helping others	.62
Being recognized by others	.67
Median	**.73**
Min	**.62**
Max	**.87**

The internal consistency coefficients are above .70 for 16 dimensions, showing an adequate reliability results of SHAPE questionnaire. There is still 4 dimensions below .70.

Study on BRAIN internal consistency

Since the new BRAIN version is adaptative, we cannot apply the methodology of Cronbach’s Alpha anymore.

Sensitivity.

Definition of sensitivity

Sensitivity, also called discrimination, refers to the degree to which a score varies with trait level, as well as the effectiveness of this item to distinguish between respondents with a high trait level and respondents with a low trait level. This property is directly related to the quality of the score as a measure of the trait.

Sensitivity provides the bottom line measure on whether a tool is highly validated or not. it reflects the ability of the questionnaire to identify the singularity of each individual. A measure is discriminant (sensitive) to the degree that it captures level of variability of interest.

The coefficient proposed by Ferguson (Delta) is a direct, non-parametric index of the degree to which an instrument distinguishes between individuals. Ferguson's Delta is the ratio of the observed between-person differences to the maximum number possible. If no differences are observed, then Delta = 0.0; if all possible between-person discriminations are made, then Delta = 1.0.

It is generally accepted that the test shows acceptable sensitivity if the Delta is over .25.

Study on SHAPE discrimination properties

The study on SHAPE discrimination functioning was conducted in April 2017.

As the results show in the Table below, SHAPE dimensions overall have very good discrimination properties, varying from .31 to .54 (mean .45).

Discrimination properties (coefficient Delta) of SHAPE personality dimensions (N=43,217).

Facets	Delta
Is assertive with others	.54
Tries to convince others	.46
Spontaneously approaches others	.55
Demonstrates diplomacy	.41
Connects emotionally	.45
Is open to other people's ideas	.39
Accepts criticism	.42
Consults others before making decisions	.42
Prefers varied tasks	.42
Is interested in abstract ideas	.40
Demonstrates inventiveness	.54
Adapts to change	.31
Organizes work methodically	.49
Pays attention to details	.50
Perseveres when confronted with obstacles	.42
Goes beyond the assigned tasks	.42
Is relaxed	.44
Focuses on the positive	.46
Controls own feelings	.50
Seeks stability	.44
Mean	**.44**
Median	**.45**
Min	**.31**
Max	**.55**

Excellent discrimination (> .50) is provided by the following dimensions: Is assertive with others, Spontaneously approaches others, Demonstrates inventiveness, Pays attention to details, Controls own feelings.

The other dimensions are also well discriminative.

Study on DRIVE discrimination properties

The study on DRIVE discrimination functioning was conducted in 2017.

As the results show in the Table below, DRIVE dimensions overall have very good discrimination properties, varying from .31 to .54 (mean .45).

Discrimination properties (coefficient Delta) of DRIVE motivation factors (N=4 550).

Facets	Delta
Create new things	.43
Excel everyday	.46
Worry about aesthetics	.45
Analyse data	.45
Meet new people	.38
Have clearly defined tasks	.41
Worry about quality	.41
Having influence	.46
Having autonomy	.51
Working in a team	.46
Having a positive impact on the world	.43
Working in a fun environment	.43
Developing in a reassuring environment	.45
Working in a disciplined manner	.53
Maintaining personal balance	.45
Receiving compensation	.45
Having attractive remuneration	.40
Achieving success regularly	.42
Helping others	.38
Being recognized by others	.32
Mean	**.43**
Median	**.44**
Min	**.32**
Max	**.53**

Excellent discrimination (> .50) is provided by the following dimensions: Having autonomy, Working in a disciplined manner.

The other dimensions are also well discriminative.

Study on BRAIN discrimination properties

To analyse the discrimination of the BRAIN questionnaire, we calculated an alpha. The higher is this alpha, the more discriminant is the item. To give some guidelines to the alphas that we obtained, Baker (2001) suggest this interpretation:
« null » when α = 0
« Very weak » when α ϵ [0,01 ; 0,34]
« Weak » when α ϵ [0,35 ; 0,64]
« Moderate » when α ϵ [0,65 ; 1,34]
« Strong » when ϵ [1,35 ; 1,69]
« Very strong » when α >1,70

Item	Alpha
$item_id_001	0.53
$item_id_002	1.72
$item_id_003	1.47
$item_id_004	1.08
$item_id_005	0.82
$item_id_006	1.60
$item_id_007	0.88
$item_id_008	0.70
$item_id_009	1.44
$item_id_010	0.72
$item_id_011	0.46
$item_id_012	1.28
$item_id_013	1.37
$item_id_014	2.52
$item_id_015	1.58
$item_id_016	1.77
$item_id_017	0.53
$item_id_018	0.95
$item_id_019	1.82
$item_id_020	2.03
$item_id_021	1.25
$item_id_022	0.82
$item_id_023	1.94
$item_id_024	1.63

$item_id_025	1.45
$item_id_026	1.98
$item_id_027	0.84
$item_id_028	0.85
$item_id_029	1.90
$item_id_030	2.46
$item_id_031	1.10
$item_id_032	0.70
$item_id_033	1.70
$item_id_034	1.65
$item_id_035	0.92
$item_id_036	1.71
$item_id_037	0.77
$item_id_038	1.42
$item_id_039	1.24
$item_id_040	1.95
$item_id_041	0.49
$item_id_042	2.27
$item_id_043	1.45
$item_id_044	1.57
$item_id_045	1.20
$item_id_046	1.01
$item_id_047	1.64
$item_id_048	1.71
$item_id_049	1.41
$item_id_050	0.55
$item_id_051	1.91
$item_id_052	1.28
$item_id_053	1.26
$item_id_054	1.46
$item_id_055	2.72
$item_id_056	1.20
$item_id_057	0.77
$item_id_058	2.76
$item_id_059	1.88
$item_id_060	2.38
$item_id_061	1.61

$item_id_062	1.23
$item_id_063	1.68
$item_id_064	0.44
$item_id_065	0.79
$item_id_066	1.29
$item_id_067	0.71
$item_id_068	0.70
$item_id_069	3.60
$item_id_070	0.90
$item_id_071	1.26
$item_id_072	1.29
$item_id_073	1.22
$item_id_074	1.84
$item_id_075	1.75
$item_id_076	1.81
Mean	**1.40**
Median	**1.39**
Min	**0.44**
Max	**3.60**

The average discrimination power of the BRAIN questionnaire is strong:
-0 item has a null discrimination.
-0 item has a very weak discrimination.
-6 items have a weak discrimination.
-31 items have a moderate discrimination.
-16 items have a strong discrimination.
-23 items have a very strong discrimination.

Why did we keep few items with a low discrimination power you may ask? Well, simply because they still bring information about very extreme levels of difficulties (very high or very low).

Fairness.

This paper discusses the question of fairness in the use of the AssessFirst solution and presents the data obtained in the analysis of the different groups studied.

AssessFirst's actions to enhance the fairness of it evaluation tool include:

#1 -Professional Orientation of the product content

The questionnaires and results of AssessFirst's solution have been specifically developed to be relevant for a professional purpose. The dimensions evaluated were chosen due to their relevance for professional efficiency.
The conclusions drawn from the use of AssessFirst are confined to this precise framework.

#2 - Validation of the questionnaires

The questions that refer to personality (SHAPE) and motivation (DRIVE) were written in such a way that they do not require specific knowledge or a certain level of education (see the results below).
Similarly, we have ensured that the answers given by the people evaluated are not affected by factors such as age or gender.

#3 - Personal Information Requested

Only necessary personal information for the proper use of AssessFirst is requested from users. There is no mention of religious, political or sexual orientation at a any time.

When it comes to age, we ask for the date of birth, to make sure that it does not affect how the questions are answered. However, this data is inaccessible to other users.

#4 - Equity of use

It is one thing for an evaluation to be designed as fair and valid and another for it to be used equitably.

To limit the subjectivity bias related to the analysis of results, AssessFirst provides users with predictive models. These are reading grids associated with a particular trade or function.

“Benchmarked” Predictive Models

This is a model based on the data collected from all the questionnaires completed for AssessFirst. These models make it possible to identify the extent to which a candidate shares the characteristics of people who are doing a job. These models make it possible to stop outside biasses from interfering, by comparing candidates with objective data.

Only dimensions that are statistically representative of the population studied are retained (with an error of 5%).

“Personalised” Predictive Models

The two main limitations of the benchmarked models is that they do not exist for every single area of work (300 are available) and that they do not reflect the specificities of the company. It is also possible to create “customised” predictive models from a qualification questionnaire for the job.

The reason for this is threefold :
- It gives a factual list of the conditions for the post
- It ensures that the dimensions present in the model are selected to aid the professional efficiency under the specific conditions chosen.
- It is important that the number of dimensions within the model are limited so that the model remains realistic.

“Talent Review”

The “Talent Review” predictive model is the most accurate technique that AssessFirst has developed to ensure maximum fairness in the evaluation process. It involves evaluating people currently performing the role and isolating the characteristics that significantly influence the performance.

These predictive models are developed by the Science and Innovation Team. Of particular attention, we are sure that gender and age do not affect the score obtained within a predictive model, or affect a candidates application.

Predictive models are widely used to appreciate the profile evaluated. To ensure diversity in these profiles, two rules are in place:
- We have limited the predictive model to 24 dimensions out of the the 45 evaluated in total. On average a predictive models includes 17 dimensions, leaving the other 28 neutral.

Therefore, two people can have the same score on a predictive model, but be can otherwise be totally different.
- On the other hand, we recommend a score of 60% for each predictive model in order to guarantee a high probability of success in a position. This leaves a 40% space between the results obtained by those evaluated and the predictive model to which they are compared.

AssessFirst's Data

The data presented in this part allows you to understand how the scores obtained in SHAPE (personality) and DRIVE (motivations) were created according to the following categories: age, gender, qualifications and also the level of responsibility.

The sample analysed includes 97,399 people who completed the questionnaire in 2017 including:
- 41280 women
56119 men
Average age - 32.4 years (average age gap - 8.7 years)

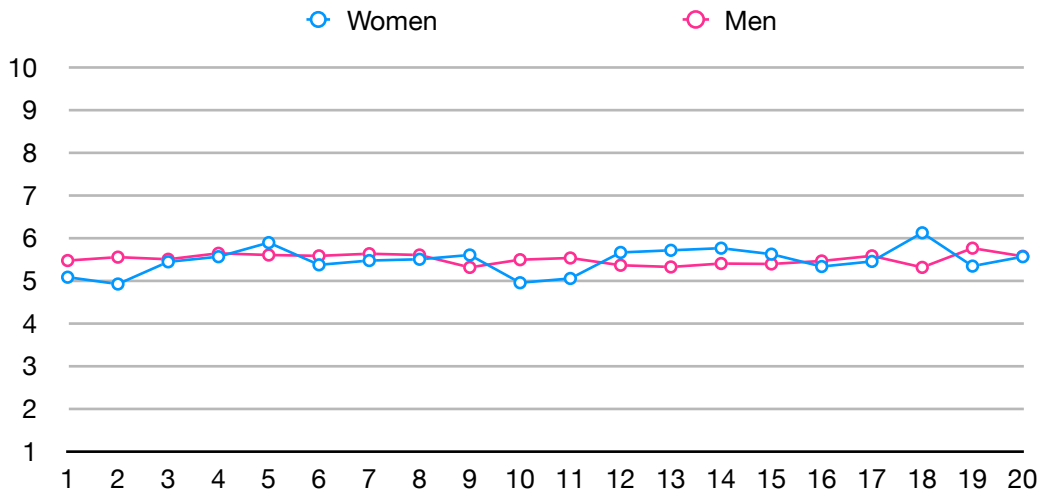
NB: To compare the results we operated the sample in to two classes from the age of 35, due to the average age indicated above.

Highest degree level
PhD - 2%
MD - 43%
BA - 34%
A-Level - 13%
Professional - 5%
No qualification - 3%

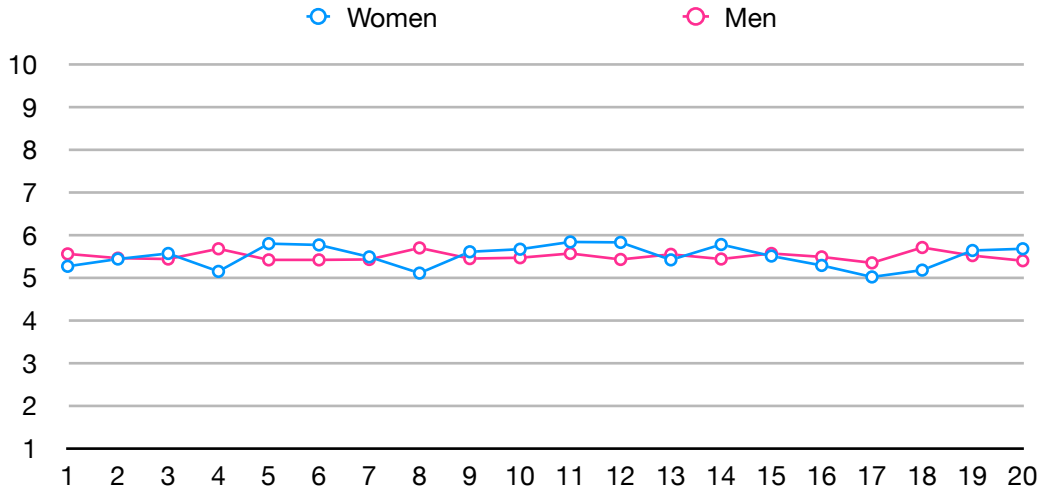
Career level
C-Level - 2%
Director - 6%
Manager - 19%
Senior - 41%
Junior - 18%
Students - 14%

Gender - Average scores (SHAPE)

There are no major differences between the results from men and women within the 20 dimensions of SHAPE. The biggest gap, with an average of 0,8 is in dimension 18 (Focus on the positive).

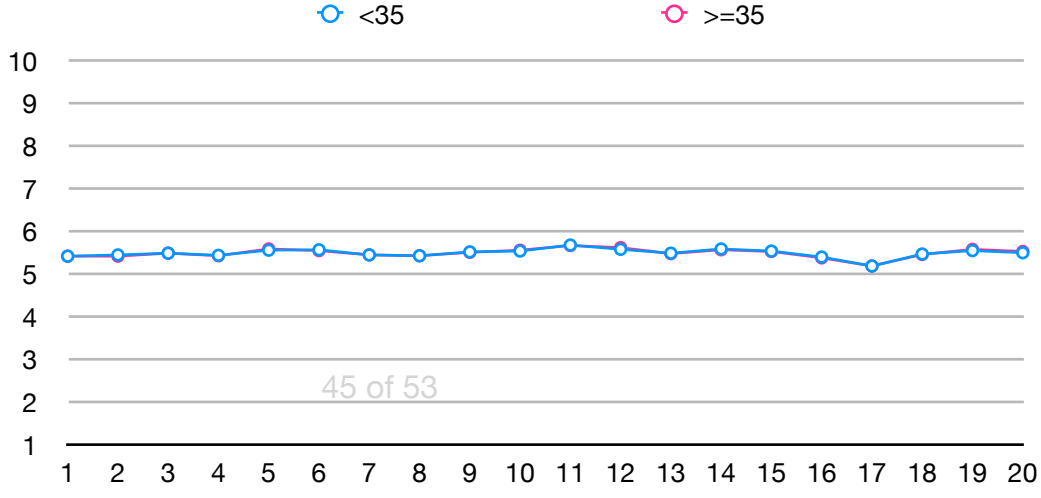


Gender - Average scores (DRIVE)



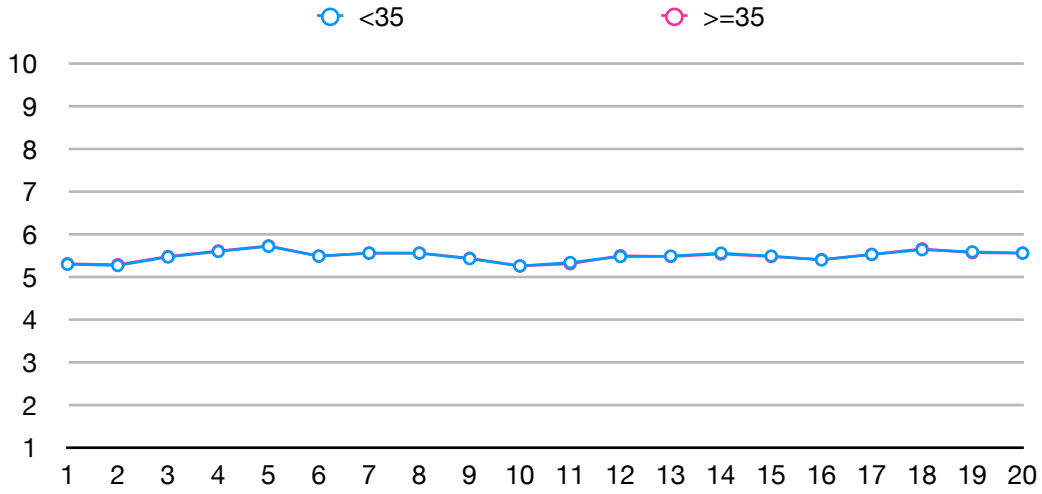
All the averages are between 5 and 6, close to the theoretical average of 5.5. Just as with SHAPE, gender does not influence the scores obtained.

Age - Average scores (SHAPE)



The scores obtained by the two groups are too close to be able to perceive any difference.

Age - Average scores (DRIVE)



Data about BRAIN (reasoning test) are separate because we did a recent upgrade of the test so the sample is (for now) 6 057 people. We'll make it evolve several times in 2021 since around 10K people are completing the test each week. The average score is 5.5 (the scale goes from 1 to 10)

Gender
Men - 28% x= 5.5
Women - 29% x= 5.5
NSP - 43% x= 5.5

Highest degree level
PhD - 2% x = 6
MD - 36% x = 6
BA - 43% x = 5.4
A-Level - 12% x = 5.2
Professional - 6% x = 4.7
None - 2% x = 4.7

Career level
C-Level - 2% x = 5,8
Director - 7% x = 5,7
Manager - 17% x = 5,6
Senior - 41% x = 5,5
Junior - 17% x = 5,7
Students - 16% x = 5

Studies about ethnicities and disabilities impact are in progress (publishing target: mid-2021).

Conclusion.

This document presents the methodology set up by AssessFirst and the results obtained to ensure maximum fairness in the evaluation of individuals.

The way AssessFirst manages data, predictive models, as well as recommendations made to users helps to ensure significant diversity within the targeted population of companies. Therefore it is possible to have tens of thousands of different result profiles and to respond favourably to job expectations.

What is more, the presentation of the results obtained by AssessFirst questionnaires shows the fairness of the evaluation, with no notable difference by age, gender, or age.

The differences that are observed, in terms of qualifications or job position can be explained by the characteristics of these variables. They do not affect variable that cannot be justified by scientific literature.

For all added information, you are welcome to contact :
Simon BARON (Chief Scientist) sbaron@assessfirst.com

References.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

An, X. & Yung, Y.-F. (2014). Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It. SAS Institute Inc. Paper SAS364-2014.

Andrews, P. H. (2006). Gender Differences in Persuasive Communication and Attribution of Success and Failure. Human Communication Research. 13 (3): 372 – 385.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. Structural Equation Modeling, 16, 397-438.

Baer, J., & Kaufman, J. C. (2008). Gender differences in creativity. The Journal of Creative Behaviour, 42(2), 75-105.

Barrick, M.R., & Mount, M.K. (1991). The Big Five personality dimensions and job performance: a meta-analysis, Personnel Psychology, 44, 1, 1-26.

Borman, W.C., Penner, L.A., & Allen, T.D. (2001). Personality predictors of citizenship performance, International Journal of Selection and Assessment, 9, 1-2, 52-69.

Bowen, C-C., Martin, B.A., & Hunt, S.T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires, International Journal of Organizational Analysis, 10, 3, 240-259.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced- choice questionnaires. Psychological Methods. Advance online publication. doi: 10.1037/a0030641.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field methods in cross-cultural psychology (pp. 137-164). Newbury Park, CA: Sage Publications.

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), International advances in self research. Greenwich, CT: Information Age Publishing.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. Psichothema, 2. 872-882.

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. International Journal of Testing, 1. 107-132.

Carducci, B.. J. (2009). The Psychology of Personality: Viewpoints, Research, and Applications. 2nd Edition. Malden: Wiley-Blackwell.

Clark, L.A., & Watson, D. (1995). Constructing validity: basic issues in objective scale development, Psychological Assessment, 7, 3, 309-319.

Cohen, R. J., & Swerdlik, M. E. (2005). Psychological testing and assessment: An introduction to tests and measurement (6th ed.). New York: McGraw-Hill.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp. 139-170).

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. Cognitive Psychology, 51, 42-100.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, Psychometrika, 16, 297-334.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Digman, J.M. (1990). Personality structure: emergence of the five-factor model, Annual Review of Psychology, 41, 417-440.

Ercikan, K. (1998). Translation effects in international assessments. International Journal of Educational Research, 29(6), 543-533.

Ercikan, K., Gierl, J. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. Applied Measurement in Education, 17(3), 301-321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), An international handbook for large-scale assessments (pp. 110-124).

Gordon, L.V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement, Journal of Applied Psychology, 35, 407-412.

Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. International Journal of Testing, 9 (2), 73-166.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. Language Testing, 20(2), 127-240.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. Applied Testing Technology, 1(1), 1-16.

Hambleton, R. K., & Lee, M. (2013). Methods of translating and adapting tests to increase cross- language validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), The Oxford handbook of child assessment (pp. 172-181). New York: Oxford University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). Adapting educational and psychological tests for cross-cultural assessment. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), Cross-cultural research methods (pp. 46-74). New York, NY; Cambridge University Press.

International Test Commission. (2005). ITC Guidelines for Translating and Adapting Tests. Bruxelles, Belgium: Author.

Jackson, D., Wroblewski, V., & Ashton, M. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? Human Performance, 13, 371–388.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. European Journal of Psychological Assessment, 15(3), 277-283.

Kankaraš, M., & Moors, G. (2010). Researching Measurement Equivalence in Cross-Cultural Studies. Psihologija. Vol. 43 (2). 121-136.

Matud, M., Rodríguez, C. C., & Grande, J. J. (2007). Gender differences in creative thinking. Personality & Individual Differences, 43(5), 1137-1147.

Pytlik Zillig, L.M., Hemenover, S.H., & Dienstbier, R.A. (2002). What do we assess when we assess a Big 5 trait? A content analysis of the affective, behavioural and cognitive processes represented in the Big 5 personality inventories, Personality and Social Psychology Bulletin, 28, 6, 847-858.

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. International Journal of Testing, 14(4), 289-312.

Rodriquez, G., Johnson, S. W., & Combs, D. C. (2001). Significant variables associated with assertiveness among Hispanic college women. Journal of Instructional Psychology, 28(3), 184-190.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? Human Resource Management Review, 16, 155–180.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. Language Testing, 20(2), 148-166.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp. 93-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

Steenkamp, J.E.M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. Journal of Consumer Research. 25, (pp. 78–90).

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), Test validity (pp. 147-169). Mahwah, NJ: Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum Publishers.

Tupes, E.C., & Christal, R.E. (1961). Recurrent personality factors based on trait rating, USAF ASD Technical Report, 61-97.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1, 89-99.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. Thousand Oaks, CA: Sage Publications.

Van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. Journal of Cross-Cultural Psychology, 31, 33-51.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing (pp. 277-308). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

Van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47(4), 263-279.

About AssessFirst

AssessFirst has developed a predictive recruitment solution allowing companies to predict how well candidates and employees will succeed and thrive in their job. The AssessFirst solution analyses data on over 5,000,000 profiles, whether candidates, employees or recruitment professionals. Today, over 3,500 companies use the AssessFirst solution to raise their performance by up to 25%, drive down their recruitment costs by 20% and reduce their employee turnover rate by 50%.

Find out more: www.assessfirst.com