MAP

Technical manual
Version 3.0

HELENE HOPPE REVALD
Director of Psychometrics, Occupational Psychologist

NANNA BAK SKYTTE
Lead Psychometrician, Occupational Psychologist

DENNIS HVASS
Psychometrician, Occupational Psychologist

Assessio

Zero Talent Waste.

Table of contents

1. Introduction & Theoretical Background

The driving force behind the development of MAP – Measuring and Assessing Individual Potential – was the need for a modern personality test with a scientific foundation, one that offers a good description of an individual's personality and with the capability to be used in the prediction of job performance with documented precision. The purpose of MAP, to provide standardized empirically documented predictions of performance to be used for decision-making, is in accordance with the international ISO 10667 standard for psychological assessments that came into effect in the fall of 2011.

The fact that human beings have different personalities and that these differences are significant for the way in which we act, is something that few people are doubting, and the scientific evidence supporting this has increased exponentially in the last decade. Personality drives human behavior, not least in the workplace. It sets the framework for an individual's strengths and weaknesses, potentials and challenges. This means that, regardless of the workplace and the tasks or the position that a person may have or apply for, personality is important for the way in which individuals view themselves and others, for how others perceive them and how they will function, thrive and perform.

The initial phase in a test development process of this kind is to define and specify the theoretical model that the instrument (measurement model) will be based upon. In the commercial test market, instruments are based on models and theories of varying scientific standards. Today, there is a broad consensus within the research community that the so-called Five-Factor Model (FFM) is the most robust and empirical model for measuring personality. This means that no matter what the purpose of identifying an individual's personality may be, e.g. selection, development or promoting self-knowledge, the empirically measurable structure is the same. The FFM is a taxonomy that postulates five broad personality dimensions: Extraversion (EX), Agreeableness (AG), Conscientiousness (CO), Emotional Stability (ES) and Openness to Experience (OP).

The FFM has a self-evident role in describing personality and is also the model with the strongest support among researchers in terms of its ability to predict behavior in the workplace. The development of the FFM started in the 1930s with the aim of investigating how many personality dimensions would be necessary to describe an "average personality" in a comprehensive manner. The researchers Gordon Allport and Harold Odbert (1936) identified around 18,000 adjectives in the English vocabulary that describe personality. By means of the at the time newly invented method of exploratory factor analysis, this number was reduced to 4,500 adjectives summarized to approximately 30 factors. There was no particular theory underlying this approach for how personality should be constructed; the factors were based entirely on everyday personality descriptions that were grouped statistically into a number of clusters.

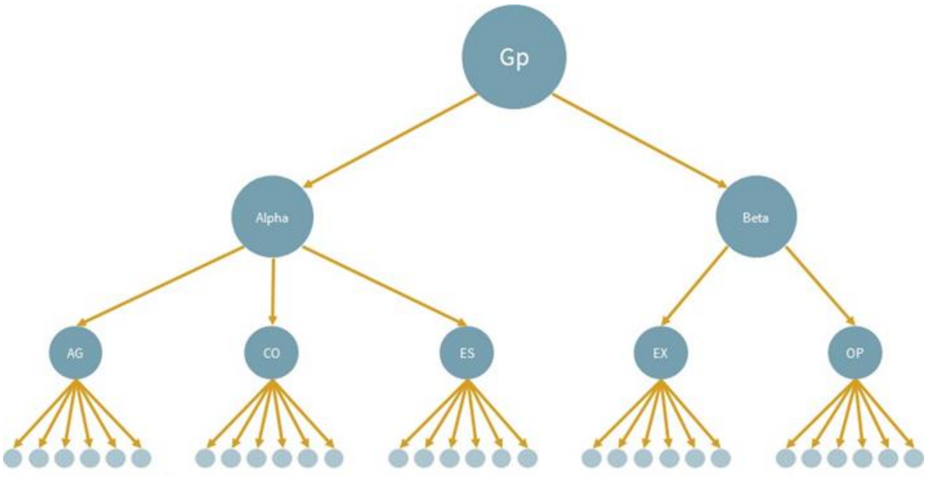
In the years that followed, research revolved around determining the number of factors and their composition, and it was not until the 1960s that the present-day FFM was formulated by Tupes and Christal (1961; 1992). This work was mainly based on extensive factor analyses of large amounts of data from the U.S. Air Force. However, in the late 1960s and the 1970s, personality research and the perception of individual differences was facing strong criticism (Mischel, 1968) leading to the FFM being somewhat forgotten. Research regarding individual differences, personality and the FFM was not taken up again until the 1980s, through longitudinal studies of personality development (Costa & McCrae, 1982). Today, the FFM is the dominating approach for measuring personality in the context of work psychology. The reason why the FFM has a special position in work psychology research is the stable empirical support showing that these factors, to varying degrees, are significantly contributing to the prediction of job performance and most other behaviors in the workplace.

Although in recent years, other models have emerged in research such as the six-factor model HEXACO (Ashton & Lee, 2008) or D-factor of personality, the FFM remains the most studied and can be considered the “gold standard” for measuring personality in research and applied settings.

The major breakthrough for personality research within work and organizational psychology came with the new methodological and statistical approach of meta-analysis. In meta- analyses, results from large number of studies are collected and summarized. This way, one obtains a more accurate estimate of the correlation between test scores and a criterion, or in this case, between measures of FFM and general job performance. The meta-analysis by Schmidt, Shaffer and Oh published in 2008 provides meta-analytic estimates of the correlations between the five personality factors and job performance. The results show that the highest correlation (p) is between the personality factor Conscientiousness (CO) and job performance. Emotional Stability (ES) has the second strongest correlation with job performance, followed by Extraversion (EX) and Agreeableness (AG), while Openness (OP) has the weakest link to job performance. Note that although some correlations may seem low, they may still have a significant financial impact on corporations and organizations in for example personnel selection (Mabon, 2005). Barrick, Mount and Judge (2001) have also shown that the importance of Conscientiousness and Emotional Stability is generalizable across occupational groups, jobs, roles, and performance criteria, while other factors may only be relevant for specific jobs or criteria. Recently, Sackett and colleagues (2021) revisited earlier meta-analytic estimates and showed that contextualized personality measures (i.e., personality tests developed specifically for a working context) showed stronger predictions of performance compared to general or overall personality measures. In addition, personality tests tend to have much less adverse impact (DEF) compared to most other methods used for selection such as job experience, GMA or work sample tests (Sackett et al., 2021).

In recent years, there has been ground-breaking research on the significance of personality in professional life and there has been great support for the possibility of describing personality in hierarchical terms, wherein the FFM represents one level. The proposed hierarchy illustrated in Figure 1 contains four levels of so-called latent constructs – abstract, psychologically meaningful entities that describe, explain and sometimes predict behaviors. The FFM-factors constitute the third level as viewed from above, the level above consists of two overall factors, to which FFM-factors contribute a varying degree and in different ways. These two so-called meta-traits are labeled Alpha and Beta, respectively, and according to Digman (1997), they can be interpreted as representations of the socialization process itself, or as personal growth. The level represented in the figure as “GP” represents the General Personality factor. This factor represents an overall psychological construct, which, according to research, would be equivalent to the so-called g-factor – the general factor of mental ability – within intelligence research. Neither the GP factor nor Alpha and Beta are operationalized within MAP. Instead, MAP measures five traits corresponding to each of the FFM-factors and the levels below, which is termed facets. Each facet is operationalized and measured on a separate scale with a unique set of items, i.e., questions, statements or adjectives, to which the respondent is asked to respond.

Figure 1. Hierarchical levels of personality.



Description of the model

Working with this type of hierarchical model has its advantages, both in terms of describing personality traits and in predicting behavior. Working on the facet level is a way of securing the construct validity, in the sense that the items used to measure a trait cover as much as possible of the domain of relevant thoughts, emotions and behaviors for the specific construct. A serious threat to the validity of personality measurement is that the psychological construct, which is supposed to be measured, is not fully covered (i.e., underrepresented). Working with a number of facets increase the likelihood of covering the content of the overall factor, since each facet is a separate indicator of the individuals' characteristics. Using facets also creates the opportunity to discover individual differences within a factor. Information on the facet level may thus offer a more detailed description of an individual or group of individuals. Two people could have the same score on one of the factors but opposite results on one or more facets. For instance, one might have high scores on facet EX1 (the first facet in Extraversion) and EX3, low scores on EX2 and EX4 and average scores on EX5, while another individual might have low scores on EX1 and EX3, high scores on EX2 and EX4, and an average score on EX5. However, facets within the same trait tend to correlate such that a person with high scores on one facet are more likely (but not guaranteed) to have high scores on other facets within the same trait.

Considering the hierarchical model, the levels above the facets may seem redundant, since the content of the facets is more specific and detailed and may thus be perceived as being more useful, detailed and thus informative. Although this might be true regarding more detailed descriptions of the individuals' personality, the further one descends into the hierarchy, the lower the reliability will be. The higher levels in a hierarchy always provide higher reliability and therefore estimate test scores with greater measurement accuracy.
Another important aspect is the possibility of generalizing, i.e. being able to draw conclusions about an individual's traits in a broader psychological sense. Since it is impractical to ask a person questions covering all possible behaviors that indicate a broad personality trait, one needs to choose certain core areas that the questions aim to cover. Hence, the questions function as indicators, from which conclusions are drawn about the personality trait as a whole. The higher up in the hierarchy, the broader generalizations are possible to make.

To summarize, the FFM model with its proposed hierarchical structure provides a clear, well-known and evidence-based framework for personality traits and facets with meaningful and measurable psychological content that is relevant for behaviors in the workplace and in general.

Initial development of MAP

Development and standardization

The development of MAP began with the definition of the theoretical model and thereby by deciding which individual differences to measure. In this work, it was decided to closely align the five major scales (traits) with the theoretical model (FFM) but allow some adaptations to a working context when operationalizing their individual components (i.e., facets).

The number of factors (five) and their general content is relatively uncontroversial from a research perspective, and there is plenty of research that links the FFM to different types of criteria such as general job performance, leadership, health, counterproductive work behavior, etc. However, the level below the five factors, thus the facet level, which entails the constructs that the facets are based upon, is not as well researched. The research literature does not offer the same clarity into the number of partial constructs below each of the five factors, their content, or how they may be measured most efficiently. Because of this, working with facets has been of a more exploratory (investigative) than confirmatory (affirmative) nature, recognizing that greater revisions might be required at this level. When used at the individual level, research shows that prediction of performance in specific jobs is greatly increased when using job-relevant facet combinations compared to the overall factors; a point that is overlooked in most (meta-analytic) studies (Barrick & Mount, 2005; Tett & Christiansen, 2007).

The facets are summed up and together they constitute the score on the scale that represents the overall factor. This "bottom-up approach" is a way of ensuring the construct validity in each part of the instrument. The final construct definitions, both for the factors and facets, their meaning, the way they are measured by scales and facets, and examples of behaviors that can be linked to high and low scores, are described in chapter 2.

Each item should be considered as an indicator of a particular aspect of the underlying construct, that is, each item should theoretically reflect the overall facet. Several indicators are required for each facet to ensure a high quality in the final measure (test score). However, according to classical test theory, it is not assumed that a certain construct is measured fully but rather that by creating multiple indicators, an individual's true score might be conveyed more precisely. True scores only exist in theory and may never be attained in practice since all measurements to a certain degree are affected by measurement errors, with error defined as either how well an item measures the same construct or by the stability of the measurement over time.

To provide a broader measure of the scales, the original theoretical model for MAP contained six facets for each of the five overall traits. The number of facets and the content of these corresponded well with the model of the personality test NEO-PI-R (Costa & McCrae, 2008). With this theoretical structure as a starting point, the work began with evaluating and revising the theoretical model and developing the measurement instrument MAP.

Based on the construct definitions of all scales, a total of 416 items with the response options Disagree, Somewhat disagree, Somewhat agree, Agree, were formulated. A team of experts (psychologists, psychometricians and test users with extensive experience in personality testing) participated in the project working with formulating and reviewing the theoretical and psychological content of each item in relation to each construct. Initially, the number of items ranged from 12 to 17 per facet. The varying number of items per facet may be attributed to the fact that certain psychological characteristics, constructs, are harder to grasp than others and therefore require a tryout of more items. Some parts of the FFM constructs are also less explored and therefore less clear in terms of what they entail and what behaviors they are likely to drive. In the end, the items underwent a linguistic control focusing on the level of readability, linguistic formulations (e.g., sentence structure, grammar), and possible difficulties of being translated and adapted into other languages.

In the spring of 2009, a questionnaire consisting of all 416 items was sent to 5 000 randomly selected people (source: SPAR register) between 18 and 65 years of age and representative of the Swedish population in terms of gender and age. In total, 569 people responded to the questionnaire after one reminder. Of these 569 respondents, 64% were women; the average age of the sample was 42 years (SD = 13). The educational level ranged from elementary school to postgraduate education. The vast majority (37%) had completed at least three years of high school and had some kind of higher education.

Once the data had been collected from the average population, the analysis was initiated. The three overall objectives entailed:

1. Ensuring that the measurement model follows the theoretical model on the FFM level.
2. Investigating the facet level by taking into account the number and distribution of scales in the measurement model across the FFM factors.
3. Given that the first and second objective was achieved, creating reliable facets (and thus indirect scales) with a maximum of 10 items per scale.

The overall model - Factors and scales

The first and perhaps most significant phase in a test development process is to ensure that the overall theoretical model, in this case the level of the five factors, is well grounded in the measurement model and that data supports the proposed model.

To test the overall model, a confirmatory factor analysis (CFA) was carried out. A confirmatory factor analysis should always be applied when there is a theoretical model to be tested because it imposes more rigorous requirements on data. The alternative approach is to apply an exploratory factor analysis (EFA) postulating no theoretical model at all to compare against. As expected, the first analysis did not meet the requirements for an acceptable adjustment and the work of revising the measuring model was initiated.

Revision of the model

As discussed above, the need for revision varied in extent and character between the five overall scales. Since weaknesses in the broader and overall model may be concealed behind lower reliability of the facets, the work started with reviewing each item in all 30 facets in relation to their “own” scale and the other four scales. The starting point of this work was the factor loadings of the facets. The goal was to create scales with high loadings on their primary scale and to ensure that the correlation between the scales was not too high while at the same time maintaining acceptable reliability levels despite the exclusion of several items.

To obtain a better fit between the measurement model and the theoretical model, the facets associated with EX (Extraversion), CO (Conscientiousness), and ES (Emotional Stability) required marginal adjustments. Items with the lowest factor loading in the respective scale were excluded. The number of excluded items varied depending on the “balance” in the loadings. In most traits, there was a clear clustering of items with a lower loading, which was therefore excluded. In the facets however, this pattern was not as clear. The items with the lowest loadings were excluded until at least 10 items remained in the facet. All items that were eligible for exclusion on these empirical grounds were examined regarding their psychological content and linked to the facet’s remaining items to ensure representativeness (content validity) in the facet. Without such a qualitative assessment, there would be a risk that the scale (or in this case the facet) might suffer from a so-called “construct underrepresentation”. This means that the items measure the construct too narrowly and thereby fail to provide indications of all the parts of the underlying psychological trait. These scales are often very homogeneous, with good internal consistency (alpha) but thus have shortcomings in terms of content validity.

The scales AG (Agreeableness) and OP (Openness) and their respective facets were identified as in need of more extensive changes, and with empirical data as a starting point, the facets were reviewed and reconstructed, taking the underlying constructs into account.

Determining the final model

At this phase in the test development process, 306 items unevenly distributed across the five traits and now 25 facets remained. The work of establishing the overall model was now considered to be completed and the remaining work focused on reducing the number of items to achieve a balanced model of 8 items per facet, thus 40 items per trait and 200 items in total. The reliability levels of the scales (which are partly a function of the number of items - the more items, the higher the reliability) also indicated that 8 items per facet would be sufficient for all facets. To achieve 8 items in each facet, items from every facet were analyzed separately using Item Response Theory (IRT). The one-parameter model was applied to each single item. In this way, items with the same difficulty levels or with greater measurement errors (residuals) were identified and then excluded after a qualitative review of the content. When the IRT approach did not point to the removal of certain items, a qualitative review of the item content was undertaken and if the content of two items overlapped, the one with the highest residual (highest error and thus lowest reliability) was excluded.

The revised model of 200 items and five facets for each of the five traits was then retested and showed a significantly better fit than the original model. This version with 200 items thus constitutes the standardized version of MAP.

Product revision (MAP 3)

In 2025, an improved version of MAP (MAP 3) was made and implemented on the platform. The revision was made to ensure that the assessment is continuously updated and maintained as job markets and societies change, research sheds light on the field, and as advanced data and new methods enable even more accurate assessments. Specifically, the revision was aimed at making the following improvements:

- Enhanced perception of relevance in a work context (i.e., contextualization).
- More accurate differentiation of candidates.
- Changes in facet names and content to meet future needs of occupational testing.
- Less risk of discrimination for different demographic groups (Adverse Impact).

An overview of the main changes is provided below in Table 1.1. Of the 25 facets, 12 were given new names, primarily to increase work relevance and better reflect the items and content of the facet in question. In addition, the contents of four facets changed significantly, and one facet was replaced to prevent overlapping with other facets and to ensure a better factor fit (ensuring that MAP is a true measure of the five factors).

The process of developing, testing, and validating the revised scales and items is elaborated in the section below on scale construction. The interpretation and content of the scales are presented in the sections below on scale definitions and content validity.

Table 1.1. Overview of major changes in MAP 3.

Scale	Previous Name	New Name	Reasons for Name Change	Content Change
EX3	Pace of Life	Work Pace	To better reflect the content of items and relevance in the work context	*No major changes*
EX4	Excitement Seeking	Risk-Taking	To better reflect the content of items and relevance in the work context	*No major changes*
AG2	Communication	Diplomacy	To better reflect the revised content of the facet and avoid confusion regarding direction of the scale	**Revised**
AG3	Altruism	Helpfulness	To make the content clearer and more separable from Compassion	*No major changes*
AG5	Affection	Conflict Aversion	To highlight that it is a new facet with new content and items	Replaced
CO1	Intensity	Accountability	To better reflect the revised content of the facet and separate it effectively from other facets	**Revised**
CO2	Diligence	Structure	To align translations in multiple languages	*No major changes*
ES1	Emotions	Unconcern	To avoid confusion regarding direction of the scale	*No major changes*
ES2	Temper	Mood Stability	To avoid confusion regarding direction of the scale	*No major changes*
ES4	Self-Control	Self-Control	–	**Revised**
ES5	Stress	Stress Tolerance	To avoid confusion regarding direction of the scale	*No major changes*
OP3	Emotional Sensitivity	Self-Reflection	To reflect the revised content of the facet and make it separable from facets within Emotional Stability	**Revised**
OP4	Experiences	Variety	To better reflect the content of items and relevance in the work context	*No major changes*

Construct definitions

The following section provides information about each constructs' position in the hierarchy of the FFM, the way it is commonly defined and measured, as well as a description of probable characteristics and behaviors for individuals with low or high scores on the specific scale. Low and high scores respectively on the scales and facets represent the opposites on the underlying dimension. This makes low and high scores more manifest and distinct on a theoretical level and more likely to represent a trait of character for an individual compared to an average score on a scale or facet. Thus, high and low scores reflect personality traits that are likely to manifest themselves more clearly in terms of their behavioral expression, compared to average scores. The section also describes which part of the overall construct each facet is intended to measure, and which behaviors that are likely to be associated with low and high scores respectively.

Extraversion (EX)

History and common definitions

The underlying construct Extraversion is found in most personality theories and is measured in nearly all available personality inventories. The actual construct, which is in fact quite broad, featuring both the need for social interaction and the level of energy, is often regarded as a not-too-difficult construct to operationalize. Extraversion is a construct that has proved to be important in some roles, occupations or behaviors in the workplace. What these have in common is their strong social component (such as leadership and service) and that they require energy and momentum (e.g., sales).

Theoretical definition

The construct forming the basis for the Extraversion scale in MAP is mainly characterized by the degree of sociability and energy directed towards the external world. Sociability includes both coping with and being interested in social interaction, as well as the need for continuous and extensive social contact with others. Both the absolute level of energy and the extent to which it is directed towards the external world are considered in this construct.

Behaviors

Individuals high in Extraversion often have a great need for and enjoy being surrounded by other people. They thrive in situations that are characterized by a high pace, and they enjoy being the center of others' attention and having the role of leaders in different groups. Some of their characteristics are talkative, enthusiastic, lively, social, light-hearted and happy. Sometimes they might also be perceived as needy, dominant, impatient, reckless, and overexcited. Because of their tendency to dominate a conversation, they might also be perceived as poor listeners. Extroverted people often give a confident impression and take up a lot of space in the social sphere. Outgoing people often enjoy excitement and seek activities and new environments that satisfy this need.

Low scores on the Extraversion scale imply a more introverted attitude towards other people and the environment, and a greater interest in one's own inner world. Focus is directed more towards one's own ideas and thoughts; stimuli from the external world, such as another person's presence or impressions, is not required. Because of this, they sometimes come across as independent and detached from what is happening in the external world. Introverted people are often perceived as less social and reserved, since they often prefer and/or need seclusion or solitude. Social engagement tires them out, while quiet environments give them strength and energy. An introverted person often prefers to work alone or in smaller intimate groups and often looks for projects that require collaboration with only a few other people and on a temporary basis. Small talk and superficial social interaction rarely interest them, which leads to them being perceived as formal, quiet, reserved and somewhat withdrawn. Introverts usually keep a low profile socially and rarely feel the need to be the focus of others´ attention. It is important to note that these individuals are not necessarily unhappy, sad or pessimistic; they just have a less exuberant expression in relation to their environment.

In MAP, the Extraversion scale consists of five facets:

- EX1. Social Need
- EX2. Social Image
- EX3. Work Pace
- EX4. Risk Taking
- EX5. Cheerfulness

Agreeableness (AG)

History and common definitions

The construct underlying the scale Agreeableness in MAP concerns how an individual interacts with other people. Although the underlying psychological construct is very broad, most measures of this trait are operationalized narrowly. Hence, scales postulating to measure Agreeableness traditionally generate a score which only indicates the degree of kindness a person expresses towards others. This aspect of an individual's interpersonal style may be relevant, but the construct is broader and thereby also needs to be operationalized with a broader set of indicators. This is a prerequisite for the test scores being valid when generalized to a broader and more significant psychological content. One consequence of this narrow operationalization is that links to relevant work-related criteria are often limited although it has shown moderate links to more specific criteria, such as service and leadership.

Theoretical definition

The Agreeableness scale offers insight into which style an individual tends to apply in their interpersonal relationships, rather than the extent of, or the focus one attributes to the social environment. The latter is captured by the Extraversion scale. An individual's social style is characterized by the extent to which the person trusts others, assuming that humans are good in general. This fundamental trust affects the interaction with others through verbal communication and body language and creates the foundation for the extent to which one radiates consideration, affection and warmth toward others. Being kind, warm and attentive are often positive traits, but the flipside of such a behavior is to show compliancy, fear of conflict, the inability to stand up for one's own opinions in relation to others as well as saying no and setting boundaries. Hence, these characteristics need to be balanced with being sincere, honest and direct in one's communication - even if one falls into disrepute, or is subjected to the anger, frustration or sadness of others.

Behaviors

People with high scores on the Agreeableness scale have a basic trust in other people. They are altruistic, caring, attentive and care about how others think and feel. They adjust their own behavior to take others' feelings into account, and they are easily affected and engaged in other people's problems and emotional states. They want and like to help and offer their support to others. They often focus on collaboration, consensus and would like to please everybody, which means that they are often perceived as helpful and tolerant. These individuals are often very comfortable and pleasant to be around, and they usually make good progress in organizations since they are often adaptable and do not question the views of others or enforce their own.

People with low scores on the Agreeableness scale are more distant, cautious and skeptical about their environment and are bound to adopt a more critical approach. They don´t attach great importance to feelings, wishes or the views of others, and rarely feel the need to adapt their own behavior to please them. They are more focused on themselves, and their approach towards others is more often characterized by competition rather than cooperation and support. These individuals might be perceived as inconsiderate, blunt, self-regarding and stubborn.

In MAP, the Agreeableness scale consists of five facets:

- AG1. Trust
- AG2. Diplomacy
- AG3. Helpfulness
- AG4. Compassion
- AG5. Conflict Aversion

Conscientiousness (CO)

History and common definitions
The trait Conscientiousness constitutes a less problematic psychological construct compared to for example Openness. Conscientiousness is well-defined and may be measured in a sound and relevant way. The width and meaning of the construct have given this trait a special position in work-related contexts in general, and in the context of psychological testing in particular.

Theoretical definition
Conscientiousness is the main personality trait of interest in most situations where performance is the focus, e.g., work, learning, education. This characteristic is constantly linked to criteria that relate to performance at work and mostly it proves itself to be the most significant personality trait. This personality trait represents the urge for achievement and includes the tendency to be organized, systematic, determined and persistent. Conscientious individuals have clear objectives, are determined and are easily able to motivate themselves. The positive aspects associated with high scores are academic and personal achievements. The negative aspects are that this may lead to excessive control, perfectionism and indecisiveness.

Behaviors
Individuals with high scores are conscientious and determined. They often possess a basic motivation for working hard, performing and achieving their set targets. They have high demands for themselves and their environment. They work methodically, systematically and in a structured manner, even when the work is monotonous, boring or requires perseverance. They are orderly, thorough and organized, which means that they are often perceived as trustworthy, conscientious and loyal members of the organization who think before making decisions. They put a lot of time and effort into preparing, organizing and scheduling, often to maintain a high and constant level of efficiency and to be able to maintain control. They may experience ambiguity and inefficiency as something troublesome and thrive in situations that are somewhat predictable. The accuracy and attention to detail can make others perceive them as controlling, overambitious or demanding, and their perseverance can lead others to the impression that they are stubborn in their way of working.

Low scores indicate that a person tends to approach their goals and commitments in a more relaxed and easygoing way. They often conceive of planning and preparation as restrictive and sometimes inhibitory and rather follow the spur of the moment, taking each day as it comes. It is easy for low scorers to delay things, since they rarely value efficiency or hard work in itself. Even when they are involved in many different things at once, their driving force is often spontaneity. This also characterizes their way of making decisions: quick and sometimes hasty, based on urges and impulses, rather than on logic and thoughtfulness. Individuals with low scores on this scale are more relaxed towards obligations and not as bothered by deadlines and commitments, or by not adhering to predetermined plans or procedures. The latter may lead them to being perceived as irresponsible, rash and unproductive, but also as flexible, spontaneous and adaptable.

In MAP, the Conscientiousness scale consists of five facets:

- CO1. Accountability
- CO2. Structure
- CO3. Ambition
- CO4. Self-Discipline
- CO5. Decision Making

Emotional Stability (ES)

History and common definitions

In an academic context, the psychological construct underlying the Emotional stability scale in MAP is called Emotional Stability, i.e., the inverse of Neuroticism. The work of measuring this trait has been going on for a long time. The first inventories of measuring neurotic tendencies were developed during World War I with the aim of assessing soldiers' ability to handle stress. Being emotionally stable has proven to be an important feature to withstand stress and strain in most professions.

Theoretical definition

The construct Emotional Stability refers both to the stability and character of an individual's overall emotional state. The construct includes both poles of the phenomenon – that is, on one hand being emotionally stable, well-adjusted and balanced, as opposed to anxiety-driven, unpredictable and unsure of oneself. A low level of Emotional Stability reflects the tendency to experience negative emotions such as fear, dejection, embarrassment, anger, guilt and disgust. Characteristics such as the ability to adapt, to resist impulses, and the extent to which one perceives and handles stress and an individual's general mood are of central importance to the construct. Note that the latter deals with the stability of moods (not being moody), rather than the type of mood (positive/negative).

Behaviors

Individuals with high scores will mostly act calmly in stressful situations. They are rarely affected by adversity and are confident and independent. One of the typical features of emotionally stable individuals is their consistent mood; they are rarely angry or provoked. Moreover, they are rarely troubled by feelings of guilt or regret over things they did in the past and are able to resist impulses and temptations, meaning that they rarely end up in unexpected situations. Individuals with extremely high scores might be perceived as insensitive or unaware of the seriousness of a situation.

Individuals with low scores on this factor tend to be anxious, worried and vulnerable to stressors. In pressured situations, people with extremely low scores might be eager, tense, restless, temperamental or easily discouraged.

In MAP, the Emotional stability scale is determined by five facets:

- ES1. Unconcern
- ES2. Mood Stability
- ES3. Confidence
- ES4. Self-Control
- ES5. Stress Tolerance

Openness (OP)

History and common definitions

The dimension Openness is the least obvious factor in the FFM hierarchy. The debate and research on the content of the construct has been going on for as long as the FFM has been applied as a framework for measuring personality. Because of this dispute, the definitions as well as the measurement have varied across different instruments, and today the definition of the construct is still unclear. It might seem incomprehensible to many, especially practitioners, that this trait often lacks relevance for performance in the workplace. In research literature, Openness rarely presents any significant links to criteria related to job performance and can generally be said to be the least important trait in terms of predicting an individual's level of job performance. However, research indicates that Openness might be better conceptualized as two factors termed sensory and epistemic Openness, the latter of which are more predictive of work performance and academic achievements (Mussel et al., 2011).

Theoretical definition

The definition of Openness in MAP is based on the original basic construct, in which Openness is an intrapsychic factor. This means that it deals with processes that occur and operate within the individual, and for the sake of the individual and their need for emotional or intellectual stimulation. Meeting this need requires an openness that allows stimuli to flow in and out. The flow that is being referred to does not necessarily take place between the individual and the environment, but in and out of the individual's inner emotional world. Stimulus can be sought from the environment but might just as well be sought for or created within the individual, without any intervention from the environment. Hence, Openness is primarily aimed at responsiveness to inner emotional experiences and openness to new experiences (be it of a sensory or intellectual nature).

Behaviors

The scale Openness in MAP involves an active imagination, aesthetics, attentiveness to one's inner emotional life, like of variety and intellectual curiosity. High scorers on Openness are curious about their own internal world and of the external world, and their lives are rich in terms of inner experiences. These individuals are likely to come up with new, unconventional and groundbreaking ideas. Their emotional world is often more intense and nuanced than people with low scores on this scale. They are often more attentive to the inner world of emotions and make room for it in their lives, for instance in decision-making (being guided by their intuition).
They often have a strong desire to try out what's new, they are open and constantly on the lookout for new experiences. These individuals seek or create situations that satisfy the need for emotional and intellectual stimulation. They often see themselves as original and artistic, while others may perceive them as eccentric, always looking for new experiences to reflect upon. Generally, they prefer the complex and often dislike traditional approaches and conservative values.

People with low scores on the Openness scale are practical and down-to-earth, focusing on what is happening here and now. They neither have a great need for, nor do they seek out new, intellectual or emotional experiences for the pure sake of the experience. They offer their own emotional world limited attention and space, both for their own sake and in relation to the external world. Their emotional reactions can, in contrast to those of high scorers, be perceived as somewhat subdued or blunt, but also as straightforward and simple. These people make logical analyses and rely on reason rather than emotions. Intuition and emotional experiences are given a limited space.
People with low scores usually prefer what's already known and teste over the unknown and uncertain. Therefore, they might be perceived as less flexible and reflecting. These individuals are more comfortable when engaging in repetitive activities and often feel satisfied when they know what to do and how to do it.

In MAP, the Openness scale consists of five facets:

- OP1. Imagination
- OP2. Aesthetics
- OP3. Self-Reflection
- OP4. Variety
- OP5. Mindset

2. Scale Definitions and Interpretations

This section contains definitions and interpretations of high and low scores for the five traits and 25 facets in MAP. Attached to each facet is an example based on the item correlating the strongest with each scale.

Extraversion

Extraversion measures sociability and the desire for social connections - the comfort level in being the center of attention, the tendency to express positive emotions and to seek stimulating experiences and a high pace of work.

Social Need
The Social Need scale measures how sociable a person is and to what extent they desire and need social interactions. High scorers enjoy both frequent and intense company of others, usually from more than a single person. They are stimulated and energized by spending time with others and are actively searching for social situations or to initiate group activities. They typically talk a lot and don't mind interacting with strangers. When they are not engaging in social interaction, they may be bored or quickly feel lonely, which can make them disturb other people to fulfill their own need. People with low scores are more reclusive and more comfortable being alone or with fewer people. Low scorers generally interact socially when it is required of them in their job but often will not take the initiative and might not enjoy it as much, as they spend a lot of energy being social and typically recharge when they are alone.

Example of an item in the facet Social Need: *I get energy from being with many other people.*

Social Image
The Social Image scale reflects a person’s tendency to be self-assertive. High scorers have a greater need for self-assertion and tend to come across as dominant and powerful. They enjoy having a prominent role and to be the focus of others’ attention. They often find it easy to gain social advantage and take a lot of space socially. High scorers can easily express themselves and enjoy taking a leading position in different groups. Low scorers on this scale have a smaller need for self-assertion. They rarely have the need to take charge, rarely dominate the social space, and rarely enjoy activities that attract the attention of others. Low scorers are more comfortable being in the background, creating and producing rather than being seen and being explicitly acknowledged. Sometimes people with low scores may be perceived as lacking in opinion or being indifferent; however, this is usually because of their limited need to draw attention to their contributions, opinions or persona.

Example of an item in the facet Social Image: *I’m often the center of a group.*

Work Pace
The Work Pace scale includes aspects such as tempo and energy. High scorers are often energetic, impatient, and are easily bored if their surroundings are not characterized by a high pace. They are often restless and have a preference for being active. High scorers often leave a vibrant impression and exude vitality but may also seem unnecessarily hectic. Low scorers generally have a more relaxed pace and prioritize time to relax. They typically prefer to have sufficient time to do their job without the need to hurry. Low scorers are more patient, which makes them appear more calm and steady but also less energetic in their expression and they do not have an equally high demand of an active and fast pace in their surroundings.

Example of an item in the facet Work Pace: *I get bored quickly if there isn’t something happening all the time.*

Risk Taking
The Risk-Taking scale concerns a person's tendency to seek excitement and unpredictability. High scorers typically enjoy taking risks and will not hesitate to take a chance if they have an opportunity to gain from it. They seek stimulation from excitement, and they will often thrive from not knowing the outcomes with certainty beforehand and doing something risky. There is a risk that these individuals are perceived as interesting but also unserious and somewhat irresponsible. Low scorers are more cautious and do not feel a need for excitement in the same way but prefer to be able to predict certain outcomes. They typically hesitate to take chances and rarely feel comfortable taking risks. They make an effort to ensure predictability of outcomes and minimize risks. They might be perceived as a bit boring by high scorers but will often also come across as serious and reliable.

Example of an item in the facet Risk-Taking: *I'm the type of person who wants to be on the safe side and I generally don't take too many risks.*

Cheerfulness
The facet Cheerfulness measures the individual's tendency to experience and express positive emotions, such as joy, happiness, satisfaction and to experience feelings of cheerfulness and being content. High scorers are generally happy and positive. They often form close connections to others more easily and may be perceived as easy to talk to by some and as a bit too much by others. They typically laugh more and express their positive emotions readily, which often means that they are perceived as cheerful, pleasant, sociable and fun. Low scorers are in general less exuberant and less lively. They may come across as formal, perhaps even a bit boring to some but more grounded to others. They may be a bit shy or reserved and may not make much of a presence in social contexts.

Example of an item in the facet Cheerfulness: I’m probably not the most bubbly type.

Agreeableness

Agreeableness assesses how a person interacts socially - level of trust, seeing others as inherently good, empathy, and the inclination to assist others and avoid conflicts.

Trust

The facet Trust measures a person's credence in others and their overall inclination to place trust in other people. People with high scores on the Trust facet believe that most people have honest and good intentions. They approach relations with trust and reliance and easily forgive others. They can be inclined to trust others blindly, making them appear naive to some, and making them more prone to be taken advantage of in some circumstances. Individuals with low scores tend to be more restrained and reserved towards others. They might be perceived as cynical, skeptical but also realistic and discerning in their approach to others. These individuals are more hesitant to place their trust in others, requiring them to prove their dependability and earn their trust over time. To them, trust is something you must earn, not assumed from the start.

Example of an item in the facet Trust: *It takes time to win my trust. (reversed)*

Diplomacy

The Diplomacy facet indicates the level of thought and concern a person puts into their communication. People with high scores in this facet tend to be diplomatic, considerate, and mindful of other people's feelings when communicating. They typically think before they speak and make an effort to be neither hurtful nor insulting in their interactions with others, sometimes running the risk of being less clear or even vague. People with low scores are often more likely to be frank and undisguised in their communication. They do not shape their way of communicating but tend to be more direct, sometimes even brutal, and run the risk of hurting other people with their bluntness. However, they may also come across as more honest and clearer when giving feedback or sharing their expectations.

Example of an item in the facet Diplomacy: *I don’t like to tell the truth if it risks hurting someone.*

Helpfulness

The Helpfulness facet measures the extent to which an individual helps others and tends to put others' needs above their own. High scorers are altruistic, attentive towards others and enjoy serving or assisting others and feel good when they can offer help. They wish to be there for others, are generous, caring, sacrificial and may have a tendency to put their own needs aside. Therefore, this focus on other's needs might make them neglect their own tasks, putting more load on themselves. People with low scores do not have the same need to be there for others and are not as attentive and aware of others’ needs. They are more self-centered and might be reluctant towards becoming involved in others’ problems. They are more inclined to see tasks as the individual's responsibility, making them prioritize completing their own tasks.

Example of an item in the facet Helpfulness: *I’m generally more attentive to other people's needs than my own.*

Compassion

The Compassion facet measures the level of tenderness, sympathy and concern for others. High scorers on this scale are considered kind and compassionate and are easily affected by the problems and needs of others. They are caring, empathetic, and tend to carry the weight of others' problems with them. People with a lower degree of Compassion may perceive them as soft and overemotional. People with low scores are more unconcerned and might appear a bit indifferent to the feelings of others. These individuals could be perceived as tough or somewhat insensitive. However, they also tend to appear calm and remain unaffected by others' emotions.

Example of an item in the facet Compassion: *I'm easily affected when my colleagues are upset.*

Conflict Aversion

The Conflict Aversion facet measures the extent to which a person tends to shy away from conflicts and a tense atmosphere. High scorers tend to give in rather than engage in a discussion to get their way. They do not necessarily have a very strong will or have any problems not getting the final say. On the contrary, they are typically content with compromises and find it more important to ease the mood in a group and do not like to stir up disagreement or share their difference of opinions. Their Conflict Avoidance might risk their contribution not being taken into account. Low scorers do not mind opposing others and find it more important to be heard than to limit disputes or avoid conflict. They are often stubborn and will fight for what they believe, which may make them come across as idealistic. Their viewpoints may be considered more often because they are not afraid to share them and hold on to them; however, that does not necessarily entail that they are more valid.

Example of an item in the facet Conflict Aversion: *I'm willing to fight for what I believe in, even if it makes me fall out with others. (reversed)*

Conscientiousness

Conscientiousness assesses how tasks are approached – the underlying drive to achieve goals and qualities like structure, persistence, thought-out decisions, and adhering to standards.

Accountability

The Accountability facet indicates the level of responsibility an individual takes accountability for at work. A person with high scores often finds themselves both competent, efficient and smart. They may have a hard time delegating tasks either because they feel responsible or because they think they might be better suited to complete the tasks themselves. They focus on the right solution to a task and act to ensure that their perspective is properly handed over, which may come across as meddlesome or even controlling to some. When troubles arise or mistakes happen, they will typically take responsibility and consider what they could have done differently to avoid the situation. Individuals with low scores do not consider themselves any more competent than others and tend not to meddle in other people's business or how they complete a task. They find it easy to delegate and do not feel particularly responsible for other people's mistakes or bad decisions. Note that this scale does not reflect whether or not a person actually is capable or suitable to perform a certain task, only whether or not this person perceives themselves as suitable and capable.

Example of an item in the facet Accountability: *I can end up taking over and doing a task myself so I know it will be done properly.*

Structure

The Structure facet reflects someone's level of orderliness and their attitude towards commitments. A person with high scores on this facet is orderly, organized, and often also perfectionistic. They might be annoyed by disorder or lack of structure, and they sometimes risk spending more time than necessary to complete a task and might also have a hard time letting go of things, especially if they consider them as imperfect or incomplete. In their endeavor to ensure structure, they might come across as a bit inflexible. Low scorers are not as organized and risk being perceived as unstructured, careless and negligent, while they might also find it easier to relate to unclear situations and instructions. They do not spend as much time organizing and structuring, navigates better in chaos, and might come across as more flexible. Low scorers rarely have trouble letting go of things, even if they are not perfect.

Example of an item in the facet Structure: *I always make sure that what I do is 100% correct, no matter how long it takes.*

Ambition

The Ambition facet reflects someone’s desire to perform, how far one is willing to go and how hard one is willing to work to achieve their goals. Individuals with high scores on this scale are hard-working, uncompromising and put a lot of effort into achieving their goals. They are ambitious, tenacious, and often have a clear direction in their work. They run the risk of being overworked and might be perceived as overly performance- and goal-oriented, and thereby hard to please. Low scorers do not have the same incentive to perform, they take each day as it comes and tend to be more relaxed when it comes to goals and performance. They are usually satisfied with their level of performance and do not find it so important to always strive for more, finding other elements in their life more important than work and performance.

Example of an item in the facet Ambition: *There are no limits to how hard I’ll work to achieve my goals.*

Self-Discipline
The Self-Discipline scale measures an individual's tendency to work hard to finish tasks, regardless of whether they are bored, distracted or find it difficult to complete them. High scores on this scale indicate that the individual has the capacity to motivate themselves and get the job done, even if it’s a monotonous, routine assignment. They often succeed in maintaining their focus despite distractions and disturbing elements. To them, the result – completing the task – is what is most important. They may have difficulties postponing things or accepting if decisions change. Low scorers may have a hard time motivating themselves to finish what they started on, especially if it is a monotonous or routine assignment that requires persistence. People with low scores are easily distracted and have difficulties staying focused on one thing at a time. On the other hand, they will not spend an excessive amount of time completing a task that may prove to be irrelevant or not worth the effort. They are generally impulsive and may be better at initiating projects than finishing them off and rarely drag tasks out.

Example of an item in the facet Self-Discipline: *I sometimes put off boring tasks.*

Decision-Making
The Decision-Making scale indicates how a person tends to make decisions - their strategy when collecting and aggregating information, ending up with a decision. High scorers are prudent and take their time to think and analyze before making decisions. They are careful, thorough and risk aversive in their search for relevant information and making decisions. This process often results in sensible and informed decisions, and they usually appear wise and trustworthy because they can usually explain how and why they made a particular decision. Sometimes, their reflections make them appear doubtful or irresolute, and they might find it difficult to make quick decisions. Low scores tend to make quick and sometimes hasty decisions, often without thinking through the consequences. They might also have difficulties gaining acceptance and respect for the decisions they make, regardless of their accuracy, since the logical rationale behind them is often missing or difficult to express. When time is of the essence, low scorers find it easier to navigate and proceed quickly to reach a decision and they will be more willing to take a risk with the potential upside that may follow.

Example of an item in the facet Decision-Making: *I’d rather act a little too slowly than risk making mistakes.*

Emotional Stability

Emotional Stability measures the extent to which someone reacts emotionally - the tendency to be calm, confident, self-controlled and optimistic, when faced with setbacks and challenges, sometimes even emotionally indifferent.

Unconcern

The Unconcern facet reflects an individual's coping strategy and how well they handle potential adverse situations. High scorers are often very optimistic and do not pay too much attention to potential risks or adversities. They are usually calm and relaxed and do not really consider what might go wrong. They have confidence in the future and their ability to deal with uncertainty, not dwelling on or regretting things that they have done. Low scorers are often worried or nervous. They may be afraid of trying out new approaches and can easily get anxious because of trivial matters or dwell on and agonize over minor decisions or mistakes. They are often tense and worry about the future. It is important to bear in mind that anxiousness also works as a driving force: being (moderately) watchful, on one’s guard and worrying about making mistakes, forgetting or omitting something, may lead to thoroughly performed work and make them take things seriously, whereas never worrying may lead to someone missing important risks.
NB: The score on this scale cannot (and should not) be used for diagnostic or any other clinical purposes.

Example of an item in the facet Unconcern: *I’m aware that I sometimes worry unnecessarily.*

Mood Stability

The Mood Stability facet measures a person’s tendency to not be irritable. High scorers have a higher tolerance level and rarely lose their temper. They generally take things lightly and it takes a lot for them to feel anger or irritation, which may make them easy for others to interact with but may also make them liable for others to lash out on. People with a low score tend to feel and express anger, frustration, disappointment, bitterness and other states related to general dissatisfaction more often than others, which makes them more likely to be perceived as irritable and moody. They often use their mood to navigate and express their opinions or set boundaries with others, which make them less likely to be exploited or taken advantage of. Note that a high score on this facet does not automatically imply that a person has a positive or optimistic appearance (this is reflected in the EX5 facet Cheerfulness), only that their proneness to irritability is low.

Example of an item in the facet Mood Stability: *My mood is so good that I’m always happy no matter what happens.*

Confidence

The facet Confidence reflects how confident individuals are in themselves, the faith they have in themselves and in their ability to handle most situations. High scorers are self-confident, have great faith in their own abilities and rarely worry about other people’s opinions. They give a more confident impression in social contexts and can easily make decisions, go for what they want and rarely question themselves or their abilities. Therefore, high scorers may at times suffer from overconfidence and rely too heavily on their own abilities and self-worth. Individuals with low scores have more doubts about their own abilities and may feel inferior to others. They often worry about what others may think, which makes them feel insecure. They get embarrassed easily and are often perceived as shy. They might have difficulties making decisions and feeling comfortable with the decisions they have already made. They may at times give an insecure impression of themselves, but they also have a tendency to be more modest and self-aware.

Example of an item in the facet Confidence: *I could use a little more self-confidence.*

Self-Control

The Self-Control facet measures the extent to which a person keeps their feelings to themselves. High scorers often take pride in acting professionally and may be hard to read. They make a great effort to appear collected and will rarely talk about their feelings or overshare. They may seem closed and sometimes even emotionally cold to others, but their self-control may mask their actual feelings, and their appearance will not necessarily reflect how they truly feel. Low scorers may have a hard time hiding how they feel. They tend to give vent to their emotions and actively share their feelings with others. They will appear very authentic when interacting with others and may overshare. Their transparency may seem overwhelming to some but can also make it easier for others to navigate accordingly. They do not necessarily experience stronger emotional reactions than high scorers but have a more immediate response to them.

Example of an item in the facet Self-Control: *I'm happy to open up about my feelings to my colleagues.*

Stress Tolerance

The Stress Tolerance facet indicates an individual's stress and strain tolerance. High scorers have a great resistance to stress and strain. They might appear calm and focused even in stressful situations, considering themselves to be capable of handling almost any kind of stressor. They often prefer a busy working environment with strict deadlines and risks being bored if that is not met. If they do feel strained, however, they often will not show, and they may be at a risk of pushing themselves too hard, not paying notice to potential stressors or strain. Low scorers are more sensitive to stress and easily get affected by various forms of strain. They may react to stress and strain in different ways and might feel incapable or unable to deal with stressful situations in an unaffected way. They often prefer not to have too strict deadlines and might have a tendency to lose focus when met with very stressful situations, often also finding it difficult to hide feeling strained. Their response to stress and stressors might let them respond to warning signs in time.
NB: Scores on this scale is not an indicator of clinical stress.

Example of an item in the facet Stress Tolerance: *I'm sometimes affected by high levels of strain and busyness.*

Openness

This trait captures overall openness to change, exploration and introspection with a preference for aesthetic solutions, variety and intellectual curiosity.

Imagination

The facet Imagination measures whether or not a person has a vivid and active imagination. High scorers are typically both visionary and creative, and they are able to find potential in unconventional ideas or solutions. They are often driven by innovation and thrive in a creative environment but may also at times promote something that is unrealistic with the risk of wasting valuable time or resources on something that will not prosper. Individuals with low scores are more down-to-earth and realistic focusing on what is feasible in the situation. They prefer to focus on the task at hand and on what is happening here and now. They may find it hard to find value in something unconventional or creative and may therefore be less likely to promote or engage in innovative or ground-breaking solutions.

Examples of an item in the facet Imagination: *I'm good at imagining things that others might find a bit unrealistic.*

Aesthetics

The facet Aesthetics reflects the amount of weight and interest an individual puts into aesthetics when evaluating, completing and designing a product or output. High scorers value beauty and appreciate great design. They typically pay a lot of attention to aesthetic details in their own and others' work and often like to work with visual aspects of a product and often do not separate function and form when making a quality assessment. Therefore, they may at times put too much time and effort into aesthetics, even when it is not necessary. To them, paying attention to details is not just about objective measures of quality but also about look and feel. People with low scores rarely show real interest in how things look because they are more interested in function than form. It is not possible to assess from a low score whether someone has bad taste; however, low scorers may not evaluate products on their visual aspects and pay little attention to the aesthetic side of things. They may be very detail-oriented from a functional perspective but might come across as sloppy because they do not pay enough attention to visual aspects of their deliveries.
Note that scores on this scale are not necessarily an indication of artistic ability or good or bad taste.

Example of an item in the facet Aesthetics: *As long as something works, it's not that important to me whether it looks nice. (reversed)*

Self-Reflection

The Self-Reflection facet indicates how receptive a person is to their own emotional state and how they tend to react in different situations. High scorers tend to notice, reflect and assign weight to their emotions and reactions. They use self-reflection as a means to understand themselves and often let their feelings guide them in decision making and self-development. Individuals with high scores often have a deep interest in understanding themselves and prioritize to be loyal to how they feel in a specific situation. Individuals with low scores are less inclined to listen to or reflect on their emotions. They do not spend much time thinking about their feelings, and do not give their emotions as much space or ascribe them as much importance and might find high scorers somewhat irrational.

Example of an item in the facet Self-Reflection: *I don't spend a lot of energy thinking about my feelings. (reversed)*

Variety

This Variety facet measures the need of a person to seek new experiences. High scorers typically seek variation in their work as a means of reaching emotional or intellectual stimulation. They are often bored if they don't continuously see or are allowed to make at least small or incremental changes in their work. High scorers are often keen on trying out new activities or tasks and may sometimes make changes for the sake of change. Individuals with low scores prefer the familiar and enjoy their routines. Change often needs to be well motivated since it is often exhausting for low scorers, and they prefer not to expose themselves to the unpredictable. Instead, they prefer to work with and improve what is already well-known and proven, tending to stick to what is familiar and already tested.

Examples of an item in the facet Variety: *I'd always prefer variety over routine.*

Mindset

The facet Mindset reflects a person's interest in and need for different forms of intellectual stimulation. High scorers generally find abstract, theoretical or complex discussions interesting and have a tendency to engage in activities that is stimulating to their intellect. They enjoy deep diving into a subject and are often open towards new ways of thinking and different perspectives. Their openness may lead them to investigate new perspectives but may also result in wasted time and energy in areas that prove less fruitful. People with low scores usually have a more limited interest in intellectual activities and are not as active in seeking out contexts or discussions simply for the sake of broadening their perspective. They are not necessarily disinterested in intellectual reasoning, but they are less likely to engage in such activities for the pure sake of it and may lose patience or fail to see the value in discussions when no clear solutions are guaranteed.
Note that scores on this scale do not indicate an individual's level of intelligence; the scores merely represent an intellectual interest or approach.

Examples of an item in the facet Mindset: *I love abstract discussions.*

3. Instructions for Use

This chapter provides guidance and advice to those who administer, interpret and provide feedback on test results generated from MAP, thus to the test administrator. In the following chapter, ‘MAP’ is referring to both full MAP and Essence unless explicitly stated or if the discussion concerns the facets, which are not included in Essence. The purpose of the following instructions, recommendations and guidelines is to create optimal conditions for the test administrator to provide the respondent with the opportunity to complete MAP in a standardized way, thereby ensuring fair and comparable results. Another aim is to create optimal conditions for the test administrator to use the results of MAP in a standardized and professional manner, for the intended purposes, and in the appropriate contexts.

Areas of use

MAP has been developed for the purpose of measuring personality according to the Five-Factor Model (FFM). This foundation makes the areas of application for MAP extensive. From research, we know that the range of areas where personality is relevant is extremely wide. Thus, MAP is developed to measure personality according to the FFM of personality and aim to provide two types of interpretations: descriptions of individuals based on their personality and to be used as a component in mechanical scoring with the purpose of predicting future performance at work.
The descriptive information provided by MAP is primarily developed to be used for onboarding, development, career guidance, teambuilding and coaching.

Administration and scoring

A test administrator may choose to start administration on site or to send an invitation to the respondent via email and thus administer MAP remotely. Administering MAP remotely makes it possible for a respondent to take the test at home or at a different location, at the same time it entails that the mode of administration is unsupervised which always means lack of control for the test administrator. This lack of control affects test administrators’ ability to ensure standardized test administration.

Regardless of administration mode, the actual test session will be identical for a respondent. The test session starts with the respondent being presented with instructions on screen and asked to provide responses according to the requirements. The set-up of the administration, interpretation, use of test scores and possible feedback is the sole responsibility of the test administrator.

Before the test session

Requirements and conditions of testing

There are several aspects requiring the test administrators’ attention and awareness before administrating MAP to a respondent. Conditions of testing and requirements for administration are therefore listed in the following along with the information that should be provided to the respondent.

Administration time

MAP is not a performance-based assessment and does not require any preparation on behalf of the respondent. However, there are conditions that must be met on behalf of the respondents for the assessment to be applicable.
The administration time, i.e. the time that a respondent has at their disposal to answer all statements, is not limited. The instructions recommend an even and steady pace when answering the statements. It is also recommended that all items are responded to on the same occasion (in a

coherent session), implying that the testing session should not take place on several, separate occasions. However, it is technically possible for the respondent to resume an interrupted or postponed session at a later point in time. Responding to the 200 items in full MAP takes approximately 20 minutes on average, and the instructions state that the respondent should set aside 20-40 minutes to complete the assessment. Responding to the 75 items in Essence takes approximately 10 minutes on average, and the instructions state that the respondent should set aside 15 minutes to complete the assessment.

As mentioned, when the test administration is completed, the test is automatically scored, calculated, and transformed into standardized test scores (hence requiring little administration time for the user). Depending on the purpose of testing, the time for both interpretation and feedback may vary.

Environment

A non-distracting testing environment is needed. Public environments, e.g., internet cafés, and public transportation, are not suitable for taking MAP. A personal computer is recommended since MAP has been visually adapted and developed for administration on a full-sized computer screen. Test-taking via tablet, smartphone or similar device is possible but may reduce the quality of the candidate experience.

A stable internet connection is needed for the full duration of the testing to ensure a valid result. Overall and regardless of the mode of administration, the test administrator is responsible for creating an accurate and friendly atmosphere; the respondent should feel safe and comfortable in the situation and given the opportunity to provide a correct picture of themselves.

The test administrator should be well acquainted with MAP – both theoretically, psychometrically and practically - and be able to convey a calm, competent and secure environment for the respondent. It is important to consider that a test situation may well be an entirely new experience for a respondent that may be in vulnerable position (as future opportunities such as job offers might depend on test results). The test administrator should spend enough time on giving the respondent a thorough introduction. He or she should have the opportunity to ask questions prior to testing, which must be answered truthfully and accurately.

It is the test administrator’s responsibility to ensure that the above requirements are fulfilled.

Target group

MAP is intended for adults, thus respondents being 18 years or older. Data for standardization, norms and psychometric evaluation is therefore always collected from individuals 18 years or older. Note that testing of minors (under the age of 18), in most geographies and legal areas require consent from the legal guardian. A test administrator may of course administer the process of consent, but it is the responsibility of the test administrator to ensure that it is done correctly and in accordance with legislation and good practice.

MAP is developed for use within the work and organizational setting, thus for selection, development and coaching etc. MAP is not intended to be used in a clinical setting or for any clinical purposes.

Computer skills

Since MAP is a web-based assessment, a certain amount of computer skills and experience of working online is required. The respondent must be able to handle the necessary technical equipment such as a mouse and/or a keyboard. Prior to the testing, it is the responsibility of the test administrator to make sure that the technical aspects do not cause any difficulties for the respondent, as this may have a negative effect on the results.

It is thus the responsibility of the test administrator to inform themselves of the respondent’s prerequisites in these matters. If there is even the slightest doubt, the respondent should be given the opportunity to demonstrate, under the supervision of the test administrator, that they possess the necessary skills to complete MAP. If a respondent is completely unfamiliar with the technical equipment (computer, tablet or smartphone) which is required to complete MAP or expresses a strong reluctance towards taking a computer-based assessment, MAP should not be administered.

Readability
Before starting the development of MAP and throughout the entire development process, it has been a true ambition to keep instructions and items as simple, clear, straightforward and as short as possible. There is nevertheless a certain demand on respondents´ levels of reading comprehension (although only a basic level of reading ability is required to understand the items and instructions for MAP). However, even if a text is classified as simple, factors such as severe reading and/or writing difficulties (dyslexia) or test completion in a non-native language may affect the understanding of content and thus the results. At present, there is no specific information regarding if, to what extent, and how MAP scores may be affected by these factors. It is therefore important that the test administrator ensures that the respondent has the reading comprehension and appropriate linguistic background required to understand the instructions and items such that the respondents can respond accurately.

Other impairments
Other impairments, including but not limited to perceptual, visual and cognitive impairments, that may have a negative effect on the test results should be identified, addressed and remedied by the test administrator before administration of MAP and any other assessment. The consequences and potential negative effects of a respondent's specific impairment when completing MAP requires investigation by the test administrator as well as deciding upon what actions are needed to most successfully accommodate and adapt the (overall) assessment process according to their specific needs.

Information to respondents before testing
The areas of application for MAP are both extensive and diverse and it is always recommended that the respondents are well informed before testing. Regardless of context, purpose of testing (development or selection), technical platform and administration mode (supervised or unsupervised), some information and instructions are crucial. Therefore, this is provided to each respondent in a standardized fashion to ensure transparency, fairness and equal treatment among respondents. This information cannot be altered or deleted.

Standardized information and instructions
Respondents are provided with standardized information and instructions on screen and prior to testing. The information and instructions provided on screen prior to starting the session include the following information:

- That MAP measures personality characteristics using statements describing different situations, behaviors, and preferences.
- That the respondents' task is to read each statement and consider to what extent it applies to them.
- That there are no "right" or "wrong" answers.
- That there will be four response options (Disagree, Slightly disagree, Slightly agree, Agree).
- That there will be two practice items, and that the responses to them will not affect results.
- That the respondent is to make sure that they can complete all statements without any distractions or interruptions.
- That there is no time limit, but it is recommended to not spend too much time on any specific statement.
- That it is possible to go back and change responses at any time during the session.

- That the respondent will answer 200 items in total if completing full MAP and 75 if completing Essence.
- That the completion time is 20-40 minutes (approximately 15 minutes for Essence).

Note that the instructions are designed to be self-instructive; all instructions needed for completion of MAP are shown at the beginning of the session. It is the test administrator's responsibility to ensure that each respondent understands how MAP is structured, how the items should be responded to, and how to complete the assessment.

Additional information

In addition to the standardized instructions (aimed at providing the most basic information and instructions on how to complete MAP), there may be other relevant information a test administrator should provide to a respondent before testing. This includes but is not limited to:

- The purpose of testing.
- What type of test MAP is in relation to why it is being used in the present context.
- How MAP will be administered and what is required for completing the test (see the previous section on requirements and conditions of testing).
- If and how the test results will be used and stored, by whom, for how long and why. Note, this does not only include the test platform but also how the test administrator/organization will use and store the data.
- That the respondent has the right to choose whether the test score is to be included as part of the information about oneself.
- Whether feedback will be provided to the respondent, when it will be provided, in what format (e.g., automatic, personal feedback, face-to-face meeting, over the phone), and what the feedback will entail.
- That the respondent should contact the test administrator in case of unexpected problems or questions arising during the test session.
- Contact details to the test administrator.

Thus, the above listed information is information that the test administrator should provide to all respondents before testing according to best practice. A test administrator may also have additional information they want to share with the respondents (for example details regarding the next step in a specific personnel selection process). Regardless, this type of information may thus not be delivered in a standardized format. If MAP is delivered through an API, the test administrator is responsible for providing the information to the respondent in a suitable way.

More information about the rights and obligations of test distributors, test administrators, and respondents are to be found in international guidelines for testing (e.g., www.intestcom.org, www.efpa.eu/professional-development, www.iso.org/standard/56436.html) and is often provided by national psychologists' associations.

During the test session

As mentioned above, the testing should take place in a calm and quiet environment, in which the respondent is not disturbed. Phones and other disturbing elements should not be present in the room. It is recommended that respondents respond to the statements at an even pace, do not spend too much time on a single item, and complete MAP in a coherent session. However, there is no time limit for completing MAP and it is possible to postpone a test session and restart at a later point in time an infinite number of times.

If MAP is administered in an unsupervised setting, it is the responsibility of the test administrator to make sure that these conditions are met during the test session and to take into account any uncertainties that might have affected results.

After the test session

MAP presupposes that a trained test administrator is responsible for the administration of MAP, the decision of whether to give feedback or not, and for providing the actual feedback to the respondent. If feedback is provided, there are several standardized written options from which the test administrator may choose from. Note however that regardless of format, the test administrator is responsible for both the decision to provide feedback or not, the choice of format for the feedback, and the content of the feedback.

It is the test administrators' responsibility to ensure that all respondents leave a test session with the feeling of having been treated fairly, having understood the purpose of the testing, and feeling that they got the opportunity to take MAP under optimal conditions. If chosen by the test administrator, they should also have received feedback on their results in a way that is perceived as fair, respectful and non-intrusive.
One of the most important aspects in this type of individual assessment is that the respondents know where to go with questions. This information should thus be unambiguous and repeated clearly to the respondent, both before, during, and after testing and after feedback has been provided.

Theoretical model, interpretation and feedback

The theoretical model underlying MAP – the Five-Factor Model – is selected, defined and developed based on the accumulated research available on measurable personality traits and the identification of personality traits that are important for predicting different types of behavior in the workplace. The meaning of the scores generated by MAP is defined by this model, by the research conducted on this model, and by the empirical support provided by the test development. Irrespective of administration purpose, the results on MAP and MAP Essence are based on the five overall traits. In addition, full MAP provides results on 25 facets (facets).

C scores and norm construction

The distribution of raw scores on each scale and facet in MAP has been transformed to the standard C scores, with a mean value of 5 and a standard deviation of 2. The method of converting and standardizing scores are elaborated below in section 7

The choice of C scores as the standard scale for the MAP is based on a comprehensible and easily communicated range (0-10) and its natural connection to the properties of the normal distribution. If test scores are normally distributed, the scale will represent the norm group accurately. Compared to other scores (such as T scores), C scores make broader classifications that is likely to convey the extent to which psychological test scores may differentiate among individuals. More finely divided scales run the risk of over-interpreting minor differences within individuals (between different scale scores) and between individuals on the same scale.

Different types of information

Information generated by psychological tests may be divided into two distinct categories. These two categories are crucial to how the results should be used, what conclusions can be drawn and what kind of decision that may be taken based on the test results. The first category of information is of descriptive character. Based on such information, an individual's traits, such as their personality, may be described according to one or several test scores that are often compared (or put in relation) to a reference group, a so-called norm group. The second category relates the test scores to a specific set of behaviors – so- called criteria – such as overall job performance or performance appraisals.

The two categories of information – the descriptive and the predictive – are often confused and it is easy to infer subjectively from a description of an individual's personality to a conclusion about an individual's future performance in a specific role and thus suitability for a specific job. However, describing an individual's strengths and weaknesses based on a personality test, or any other assessment tool, does not automatically mean that the individual is suitable or unsuitable for a specific task or position. Descriptive information is of course useful for other purposes, such as self-awareness and individual development.

Descriptive information

The vast majority of tests and other assessment methods generate data of descriptive character. An individual's traits are expressed in test scores determined in relation to a relevant comparison group, a so-called norm group. A norm group may be described as a group of individuals who have also taken the assessment under equivalent conditions, to which an individuals' score is compared.
It is important that the norm group is relevant for comparison such that the normative score is meaningful and comprehensible for the test administrator and the respondents. To the respondent, the comparison with the corresponding description will be comprehensible when it's done in relation to what they may think of as "most others". An individual's relative image of their personality is usually based on comparisons with most other people and not on comparisons with, for example, a specific occupational group or a specific organization. To describe a person's basic traits in relation to a specific and perhaps extreme group in terms of personality (for example a specific occupational group) may create confusion among respondents, as they are unable to recognize themselves.

Individuals who have already been tested or undergone psychological assessments on previous occasions may experience conflicting results because of various norm groups being applied for different assessments. Even if the norm group is explicit, which is not always the case, it is difficult for the respondents to keep their traits in mind throughout the assessment process, which might include multiple methods and comparison groups. The relevance of comparing the respondent against a specific group may often be questioned. It is important that the respondent, regardless of the decision or the measure, perceives the testing and feedback as being understandable, relevant and meaningful, not only in the current situation but also in the future. For example, it is questionable to give a respondent the feedback that they are introverted, based on the comparison against a norm group representing a specific group (e.g., an occupational group) having high scores (and often a skewed distribution) on Extraversion, if he or she is actually outgoing, compared to a more general population.

General principles for interpretation and feedback

MAP may be applied in many contexts and the context to some extent determines the way in which the results should and can be used adequately. For example, in a selection context the empirical correlation between scales and job performance should be the main focus. In a development context, detailed, individual feedback is likely the most relevant aspect. Regardless of purpose, the results are interpreted, and the respondent is likely to receive some form of feedback. The following recommendations concerning interpretation and feedback are general in nature and may not apply to all contexts.

The meaning of test scores

The quality and empirical foundation for MAP is not the only aspect affecting the quality of the interpretation of results. The test administrator's understanding of the theoretical model, the underlying psychological constructs and the knowledge about the data forming the basis of the test scores is also relevant for the understanding of test results and thus giving feedback. Prior to giving feedback to a respondent, the test administrator should engage in:

1. Getting to know the core of every psychological construct as defined in this manual (i.e., the construct underlying each scale), without ascribing subjective meanings or associations from one's own behavior to them outside of what is described as the core of the construct. The possibility of making generalizations based on test scores is given by the definitions.
2. Evaluating the results of the scales. The order of the scales (EX, AG, CO, ES, OP) have no meaning in itself. The scales may be interpreted in whatever order suits the test administrator or according to what suits the purpose of use. In general, the following order and grouping is recommended for the interpretation and feedback of the scales:
    - EX and AG reveal the way in which an individual acts and functions in the interpersonal sphere (between individuals) in terms of outgoing energy and what style one tends to have in social interactions.
    - CO and OP are related to the way in which the individual's intrapersonal sphere (within the individual) operates in terms of new emotional experiences, and whether or not they are conscientious and focused on performance.
    - ES is suitable to interpret lastly because it indicates an individual's general emotional reactions and they handle insecurity, adversity, stress, or strain.
3. If using full MAP, evaluating the results on the facets. Examine questions such as: Which facets seem to be the driving forces for the results on the trait? Are average scores on traits reflecting a uniform or a varied profile of results on the facets? Regarding high and low scores on this scale: Are there any facets whose scores are going in the opposite direction, and so on. Keep in mind that facets within the same trait tend to correlate, so the facets will tend to follow the same direction as the result on overall trait (although some deviation is not unusual).
   Unusual patterns of scores on the facets might be difficult to understand, especially for new users, if too much emphasis is put on the adjectives chosen to label them. Keep in mind that the labels can never reflect a complete and perfect description of a psychological construct or trait, and they cannot describe how the construct differs from other constructs (when considered in isolation).
4. Understanding what is actually measured is as important as knowing what is not measured in each scale. One should also keep in mind not to over-interpret differences connected to point 2 and 3 above and to consider the aspect of measurement error throughout the process.

Average scores

Sometimes average scores are perceived as more difficult to interpret and to give feedback on, compared to high or low scores. This might especially be the case when a respondent has many average results compared to the number of high or low results. In the case of most scales and facets, one often feels greater certainty interpreting combinations of more extreme results and significantly different levels of test scores. When the score falls within the average range, we tend to be less certain of how to interpret and thus give feedback. In MAP, most scores (54 %), will fall between C scores 4 and 6. A respondent with an average result are like most others in the norm group regarding this trait. Average results reveal that the particular trait is unlikely to be a strong characteristic for the respondent, thus, it will most likely not be perceived as one of their prominent traits. The expression of an “average trait” is likely to be more moderate than the expression of a trait on which the respondent has gained high or low scores. It is also important to note that the width of underlying constructs makes average results on traits represent a broader psychological meaning than average results on facets.

Feedback

In line with current best practices and guidelines, the overall recommendation is to provide respondents with some kind of feedback regarding their test results on MAP. Respondents should also be provided with the opportunity to ask questions regarding their test results. The feedback, however, may be delivered in different ways. The appropriate form of feedback in a specific context is dependent upon several factors, for example the type of assessment (thus the actual content), the number of respondents which is nested in a financial aspect (face-to-face feedback is time consuming and thus expensive while written standardized feedback is more cost effective), and the availability of the respondent (respondents may be at remote locations).

One of the most common ways to provide feedback on test results in practice is to provide the respondent with a standardized written feedback report generated by the system. This approach should be accompanied by a clear offer to the respondents to inquire about their results (for instance, via phone or email). The second most common approach is to provide verbal feedback (in a face-to-face or virtual meeting or via the phone). Usually, the results are reviewed and discussed, and the respondent is provided with the opportunity to comment on the results. Often the test administrator will use the standardized descriptions generated by the platform as the basis for feedback. In general, the standardized feedback generated by the platform and intended for respondents does not require any additional personal feedback from the test administrator.

If feedback on test results is carried out verbally, it is recommended to start with an initial conversation regarding how the respondent experienced the test session, how it was perceived, whether the respondent felt that they were given the opportunity to provide an accurate picture of themselves, if they understood the instructions, if they were able to answer all statements with comfort and if the respondent would like to comment on any circumstances that they think may have affected results. This should take place before feedback is given on the results. After this, it is appropriate to describe the structure, the purpose and the length of the feedback session, and to encourage questions and reflections during the entire feedback session. Stress that no test (assessment) results are exact, they are always affected by measurement error, and that adjectives used to label the scales may never fully reflect the underlying psychological construct. Encourage the respondent to ask questions throughout the feedback conversation and as a test administrator, explore the hypotheses regarding the respondent’s personality that are suggested by the results.

Start the feedback by explaining why it is relevant to measure personality in the given context. Describe the overall structure of scales and give an overview of how the results will be reviewed. Then, start with EX and AG. Describe their position in the interpersonal sphere, their main content, and describe the facets of the respective scale. Proceed to the respondent’s score on these traits and facets. Proceed in the same manner with CO and OP by describing their intrapersonal characteristics and the main content of the scales, followed by the respondent’s result. Conclude with the ES scale.

4. Scale construction

As described earlier, MAP is a contextualized measure of the Big Five personality traits. This section outlines the process of item development, data collection, item selection, and contextualization for the most recent version of MAP (MAP 3).

Item development

Based on the theory and research presented above, a large pool of items for each scale were produced by our internal team of psychologists and psychometricians. In this process, careful consideration was given to the relevance and appropriateness of items in a working context (i.e., items of a personal or intimate character were omitted). To ensure content validity, items were written based on the vast amount of research available on the Big Five (cited in earlier chapters).

The best items (in terms of wording as well as content) for each scale were then tested and reviewed in successive fashion, with each scale requiring a different number of iterations to reach the final scale. In total, 775 items were submitted to testing, of which 200 (26 %) were approved.

Data collection

Data on items submitted to testing was collected from February 2024 to March 2025. For all items, a minimum of 500 cases were collected prior to statistical analyses and psychometric scale validation. Demographic criteria were similar to those of the norm group, i.e., people aged 18-70 years of age who completed the test in a high-stake setting (predominantly selection) and in their native language.

Item selection

Upon data collection, items were subjected to a series of statistical analyses and reviewed by at least two reviewers from our internal team of psychologists and psychometricians. At the first stage, items were evaluated with respect to internal consistency (by means of Cronbach's alpha and corrected item-total correlations) as well as response distributions to detect lack of discriminatory power or socially desirable responding. Items were screened out if less than 10 % or more than 90 % of answers were in the same direction (i.e., response options 1 and 2 or 3 and 4 combined).

Items were selected based on the corrected item-total correlations using a top-down procedure, which has proven one of the best methods for item selection (Zijlmans et al., 2019). In this procedure, all items are included at the first step. Based on the corrected item-total correlations, weakly correlating items are then removed successively, until the scale can no longer be improved (i.e., the point at which the exclusion of any item would not enhance the internal consistency of the scale). Although higher values were desired for the final scales, a guiding cut-off for the correlation was set at .20 during development to avoid discarding items with relevant content and sufficient psychometric quality.

Final scale validation

Following the various number of development iterations, each of the preliminary scales was then subjected to an extensive psychometric validation procedure assessing six different aspects:

1) Consistency
2) Unidimensionality
3) Local independence
4) Item invariance (no Differential Item Functioning, DIF)
5) Scale characteristics
6) Group differences & Adverse Impact

Each of these aspects is elaborated in the sections below on construct validity. In addition to the psychometric validation, the final scales were also examined for qualitative aspects such as readability, comprehensibility (negations), grammatical redundancy, and balancing the number of positively and negatively worded (reversed keyed) items.

In total, 200 items were included with each scale consisting of 8 items.

5. Validity

Despite a clear definition of validity, it is a topic of debate among international researchers and experts, how many types of validity there are, and which research methods are most suitable for shedding light on what. This is mainly due to the fact that, in practice, it can be difficult to determine the type of validity a given study relates to. However, there is a growing consensus that validity is a unitary concept, which can be documented by various forms of statistical and empirical studies.
In the following, validity is categorized and divided into face, content, construct, and criterion validity in accordance with the EFPA test review model (EFPA, 2013).

Face validity

Face validity concerns the extent to which test users and test subjects perceive the questionnaire and test results as relevant, comprehensive, and reflective of reality. Face validity is thus about whether a test comes across as credible to the test person, which is important to ensure that a test person is sufficiently motivated to participate in the test and accept the conclusions drawn from it. It is also about recognizability of test results to both the test person and to others.

To ensure face validity, two key aspects were considered during the development process. First, items with a high degree of transparency as to what is being measured were preferred over dubious items with less recognizable relevance (which is often the consequence of using a solely data-driven item selection process such as empirical keying). The use of mainly transparent items has the advantage of providing a clear link between the test subject’s answers to the individual items and the final scores derived from them. Hence, the likelihood of test subjects understanding, recognizing, and accepting test results is increased, which is crucial for self-awareness and general usage of the test for work-related purposes (be it recruitment or personal development). Although this transparency can make items more susceptible to faking or socially desirable responding (i.e., impression management), statistical analyses conducted during item development suggest that this is not a major cause of concern.

Second, items were carefully written to reflect behavior in a working context in keeping with the principle of contextualization, e.g., being social with or helping out colleagues (as opposed to friends or family). In addition to the items, the aspect of contextualization is also reflected in the naming of scales, e.g., Work Pace and Risk-Taking as opposed to Activity and Excitement-seeking, as they are named in the original FFM framework.

Item examples for each scale can be found in the section above on scale definitions and interpretations.

Content validity

Content validity concerns whether test items and scales constitute a relevant, representative sample of the aspects that define the theoretical concept (domain) being measured.

With 100 years of research and a vast number of scientific publications, the NEO-PI-R (Costa & McCrae, 2008) represents the basis from which to select and organize each of the scales (traits) and facets (facets). However, as NEO-PI-R was invented for a range of purposes, including clinical ones, a number of alterations were made to the original model. As such, MAP represents a contextualized measure of personality as expressed in a working context. In the following, we present these adaptations for each of the five traits alongside the theoretical linkages between MAP scales and NEO-PI-R facets (presented in Tables 5.1-5.5).

Extraversion (EX)

Conceptually, the Extraversion facets in MAP are closely linked to the facets in NEO-PI-R. The facet of Activity was renamed Work Pace to better reflect how this behavior plays out in a working context (e.g., being energetic and preferring to stay busy). Similarly, Excitement-seeking was renamed Risk-Taking to capture the tendency and willingness to take risks within the realms of work as opposed to non-work-related behaviors such as gambling or seeking out dangerous activities.
The sixth NEO-PI-R facet of Warmth (related to interpersonal intimacy) was removed, as this scale shared too much conceptual overlap with the trait of Agreeableness as confirmed by empirical analyses from the initial construction of MAP.

Table 5.1. Comparison of Extraversion facets in MAP and NEO-PI-R.

Extraversion		
No	MAP	NEO-PI-R
EX1	Social Need	Gregariousness
EX2	Social Image	Dominance
EX3	Work Pace	Activity
EX4	Risk-Taking	Excitement-seeking
EX5	Cheerfulness	Positive emotions

Agreeableness (AG)

Regarding Agreeableness, two major adaptations were made. First, the NEO-PI-R facet of Straightforwardness was operationalized to capture aspects of Diplomacy, with low scores reflecting a tendency to be very honest and direct when giving feedback to others (to the point of being blunt) and high scores reflecting a tendency to adjust one’s communication or withholding criticism (at the expense of honesty). This adaptation was made, as a measure of straightforwardness (being sincere vs. manipulative) revealed inferior correlations with other AG scales and a stronger fit to other traits (CO, ES).

Second, the NEO-PI-R facet of Modesty was dropped from this trait, as the contextualized version (being humble regarding one’s abilities) was inseparable from low scores on Accountability (in CO) and Confidence (in ES) (interestingly, both of these facets are dropped from AG and placed in the sixth trait of Honesty-Humility in the HEXACO model; Ashton & Lee, 2008).

Regarding Compliance, we chose the name Conflict Aversion to separate clearly from aspects of complying with rules, structures, and protocols, which is captured by Structure (in CO).

Table 5.2. Comparison of Agreeableness facets in MAP and NEO-PI-R.

Agreeableness		
No	MAP	NEO-PI-R
AG1	Trust	Trust
AG2	Diplomacy	Straightforwardness
AG3	Helpfulness	Helpfulness
AG4	Compassion	Tender-mindedness
AG5	Conflict Aversion	Compliance

Conscientiousness (CO)
Most notably, the NEO-PI-R facets of Order and Dutifulness were combined into a single scale (termed Structure) as these scales had very high factor loadings on the overall CO scale and showed substantial correlations with one another in the initial stages of constructing MAP. Statistical analyses confirmed that these facets can be meaningfully combined into a single scale, as items relating to structure, rigidity, perfectionism and adhering to rules, obligations and principles showed several cross-loadings when modeled on two separate factors. Furthermore, upon careful selection and revision of items, the scree plot from the factor analysis supported the extraction of just a single factor. Although representing a mix of facets, the predominant weight was put on structure and perfectionism, as being dutiful in terms of complying with social standards and expectations are better captured elsewhere (e.g. in the Match-V scale of Conformity).

Table 5.3. Comparison of Conscientiousness facets in MAP and NEO-PI-R.

Conscientiousness		
No	MAP facet	NEO-PI-R
CO1	Accountability	Competence
CO2	Structure	Order + Dutifulness
CO3	Ambition	Achievement striving
CO4	Self-Discipline	Self-discipline
CO5	Decision-Making	Deliberation

Emotional Stability (ES)
To capture Emotional Stability, all of the initial NEO-PI-R scales within Neuroticism (variously termed “Emotionality” or “Emotional reactions”) were inversed such that high scores on MAP correspond to low scores on the corresponding domain.

For Self-Control, an adaption was made to focus on controlling or hiding one’s emotional expressions (as opposed to being more authentic and transparent). This operationalization differs slightly from the corresponding NEO-PI-R facet of Impulsiveness defined as a lack of ability to control urges or cravings (in terms of food and immediate needs). However, in a working context, this would entail being easily distractable and struggling to remain focused on the task at hand, thus making the scale redundant with Self-discipline (within CO). Furthermore, aspects of impulsivity related to risk raking or spontaneous decisions are better captured by Risk-Taking (within EX) and Decision-Making (within CO), respectively.

The final NEO-PI-R facet of Depression was left out for several reasons. First, the content of the scale is very clinical in nature, thus raising both practical, legal, and ethical concerns. Second, non-clinical aspects relating to the perception of negative feelings are captured in low Cheerfulness (within EX). Finally, the aspect of having low expectations for the future (i.e., pessimism) was deemed inseparable from low Unconcern (as confirmed by strong factor loadings for these items on this scale).

Table 5.4. Comparison of Emotional Stability facets in MAP and NEO-PI-R.

Emotional Stability		
No	MAP facet	NEO-PI-R
ES1	Unconcern	Anxiety (R)
ES2	Mood Stability	Hostility (R)
ES3	Confidence	Self-consciousness (R)
ES4	Self-Control	Impulsiveness (R)
ES5	Stress Tolerance	Vulnerability (R)

Openness (OP)
In the FFM, Openness is by far the most broad and heterogeneous trait, which is reflected in historical controversies surrounding its definition, structure, and even name. Owing to its complex nature, a range of names such as Interest, Intellect (or Intellectance), Independence, Open-mindedness, Creativity or Curiosity. In addition, several meta-analyses have shown Openness to be the trait least predictive of job performance (Mussel et al., 2011), although Sackett et al. (2021) showed higher validities for contextualized vs. general personality measures (for all traits, including Openness).

In an attempt to resolve these controversies, Mussel et al. (2011) suggested a two-factor structure of the construct separating perceptual Openness (Fantasy, Aesthetics, Feelings) from epistemic Openness (Actions, Ideas, Values), the latter of which predicted job performance and academic success better than did the first factor.

For these reasons, a number of adaptations were made to the first three facets to enhance contextualization and increase the likelihood of predicting work-related behavior. First, Fantasy was relabeled Imagination measuring the tendency to be creative, visionary, and “thinking outside the box” (vs. applying a more practical and realistic approach in dealing with various tasks).
Second, Aesthetics was measured using items related to the workplace or work tasks. i.e. a dimension ranging from utility or functionality at the low end to aesthetics or “look and feel” at the high end (e.g., prioritizing functionality and presentability equally, if not higher).
Third, Self-Reflection emphasizes aspects of self-reflection (or self-awareness) and how one deals with emotions in a working context (e.g., how they affect decisions or work processes).

Regarding the final three facets, the operationalization of Variety and Mindset was quite close to their corresponding NEO-PI-R facets emphasizing actions (trying something new, aversion to routine) and ideas (pondering abstract ideas and engaging in theoretical or philosophical discussion), respectively.

Finally, the NEO-PI-R facet of Values was removed from MAP, as the scale content showed a considerable overlap with the Match-V scale of Idealism (sharing aspects of tolerance and open-mindedness). In addition, the facet of Values includes questions of an ideological or slightly political character (e.g., whether someone respects foreign traditions or religious authorities), which were deemed inappropriate to use in a working context (especially for selection).

Table 5.5. Comparison of Openness facets in MAP and NEO-PI-R.

Openness		
No	MAP facet	NEO-PI-R
OP1	Imagination	Fantasy
OP2	Aesthetics	Aesthetics
OP3	Self-Reflection	Feelings
OP4	Variety	Actions
OP5	Mindset	Ideas

Construct validity

Construct validity concerns the agreement between test results and prior theoretical knowledge of the construct being measured. As this validity aspect is quite broad, it is further divided below according to the final scale validation mentioned previously.

Consistency

First, internal consistency was measured using Cronbach's alpha. However, the alpha coefficient suffers the well-known drawback that it relies not only on the correlation between items and the scale but is also affected by the number of items in the scale (Taber, 2018). Furthermore, studies show that too high levels of alpha can be undesirable and often occur because of redundancy among items (Tavakol & Dennick, 2011).

Therefore, the alpha coefficient was supplemented by corrected item-total (item-rest)correlations, in which each of the items is correlated with the rest score (i.e., the scale score minus the score for the item in question). The higher the correlations, the higher the internal consistency of the scale. Also, consistency was evaluated by ensuring that all inter-item correlations were positive (as recommended by Streiner and Kottner, 2014).

Cronbach's alphas, average item-rest and inter-item correlations for each scale are listed below in Table 5.6. All scales showed sufficient (.70) or good (.80) levels of consistency as defined by the EFPA test review model (EFPA, 2013). Alphas ranged from .70 to .86 with an average of .76 and a median of .75. Across the five traits, median alphas were .79 for EX, .77 for AG, .74 for CO, .74 for ES, and .79 for OP (please note that when summed into a composite score, all of these scales would exceed the .80 or .90 mark recommended for selection). In addition, item-rest correlations were all above the desired value of .30 (He & Wang, 2015; Shen et al., 2018), and all inter-item correlations were positive with a scale average ranging from .22 to .40.

Table 5.6. Internal consistency of MAP scales.

No	Scale	Alpha	Item-rest cor.	Inter-item cor.
EX1	Social Need	.79	.50	.32
EX2	Social Image	.80	.51	.33
EX3	Work Pace	.76	.47	.26
EX4	Risk-Taking	.86	.60	.40
EX5	Cheerfulness	.74	.43	.26
AG1	Trust	.80	.50	.34
AG2	Diplomacy	.77	.47	.31
AG3	Helpfulness	.74	.44	.28
AG4	Compassion	.78	.49	.31
AG5	Conflict Aversion	.70	.40	.24
CO1	Accountability	.70	.39	.25
CO2	Structure	.71	.40	.25
CO3	Ambition	.74	.43	.29
CO4	Self-Discipline	.75	.44	.27
CO5	Decision-Making	.74	.44	.27
ES1	Unconcern	.76	.45	.29
ES2	Mood Stability	.72	.41	.23
ES3	Confidence	.74	.44	.27
ES4	Self-Control	.75	.45	.27
ES5	Stress Tolerance	.70	.39	.22

OP1	Imagination	.80	.51	.33
OP2	Aesthetics	.79	.50	.34
OP3	Self-Reflection	.81	.53	.34
OP4	Variety	.73	.42	.27
OP5	Mindset	.74	.43	.24
Mean		**.76**	**.46**	**.29**
Median		**.75**	**.44**	**.27**

Unidimensionality

Unidimensionality of the scales was investigated using an exploratory factor analysis (Principal Components Analysis, PCA) requiring the extraction of just a single factor. For all scales, a visual scree plot supported a single factor solution, which was confirmed by strong factor loadings, i.e., correlations between scores on the individual items and the score on the extracted factor.
An example Scree plot from the scale Social Need is shown below in Figure 5.1.

Table 5.7 displays the Eigenvalue and the amount of explained variance (%) for the single factor solution alongside the items' average factor loading for each scale. Across scales, the variance accounted for by any single factor ranged from 32.8 % to 50.5 %. In addition, all average factor loadings exceeded the .40 mark recommended by Howard (2016), ranging from .56 to .70 with an average of .61 and a median of .60 across scales. In sum, these analyses support the unidimensional nature of the constructed scales.

Figure 5.1. Example Scree plot from Social Need.

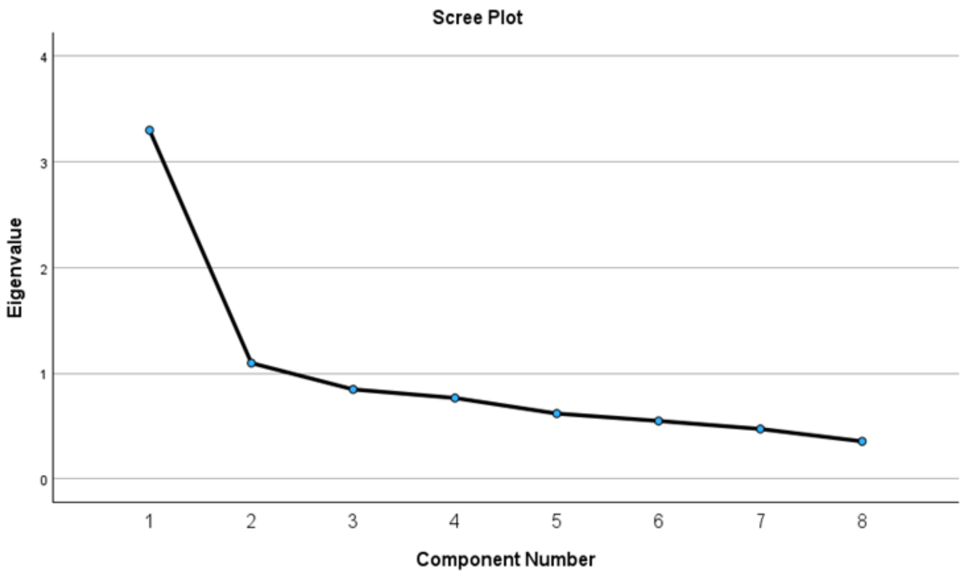


Table 5.7. Unidimensionality of MAP scales.

No	Scale	Factor loading	Eigenvalue	Expl. Var.
EX1	Social Need	.64	3.30	41.2
EX2	Social Image	.65	3.40	42.5
EX3	Work Pace	.61	3.11	38.9
EX4	Risk-Taking	.70	4.04	50.5
EX5	Cheerfulness	.59	2.87	35.9

AG1	Trust	.64	3.35	41.9
AG2	Diplomacy	.62	3.13	39.2
AG3	Helpfulness	.60	2.90	36.3
AG4	Compassion	.63	3.23	40.3
AG5	Conflict Aversion	.57	2.68	33.5
CO1	Accountability	.56	2.62	32.8
CO2	Structure	.57	2.70	33.8
CO3	Ambition	.59	2.87	35.9
CO4	Self-Discipline	.60	2.91	36.4
CO5	Decision-Making	.60	2.88	36.0
ES1	Unconcern	.60	2.98	37.3
ES2	Mood Stability	.58	2.76	34.5
ES3	Confidence	.59	2.97	37.1
ES4	Self-Control	.60	3.04	38.0
ES5	Stress Tolerance	.57	2.64	33.0
OP1	Imagination	.64	3.36	42.0
OP2	Aesthetics	.64	3.31	41.3
OP3	Self-Reflection	.65	3.56	44.5
OP4	Variety	.59	2.82	35.3
OP5	Mindset	.59	2.92	36.5
Mean		**.61**	**3.05**	**38.2**
Median		**.60**	**2.97**	**37.1**

Local independence
As stated above, high inter-item correlations were desired to improve the internal consistency of scales. However, too high values can reflect local dependence, i.e., that the response to one item depends on the response to another item, which inflates overall estimates of scale reliability (Marais, 2012). Ultimately, it can also jeopardize the candidate experience because candidates need to spend more time completing the assessment and experience what feels like answering the same item twice.

Local independence was inspected using partial correlations, where each of the items are correlated whilst partialing out the effect of the scale score (i.e., the sum of items). Hence, partial correlations are expected to be weaker than inter-item correlations and close to zero or negative (van Bork et al., 2018).
As the magnitude of partial correlations depends on both the number of items in the scale, the sample size, and the number of response options, it is difficult to set fixed criteria as to what constitutes a too high partial correlation when considered in isolation. Therefore, the largest observed partial correlation was subtracted from the average partial correlation to obtain a relative estimate of local independence similar to the Q3* measure used within Rasch and IRT models (Christensen et al., 2017).

The average and maximum partial correlations as well as the maximum differences for each of the scales are displayed in Table 5.8. Across scales, the difference between the maximum and average partial correlation ranged from .13 to .35 showing sufficient local independence for each of the scales as suggested by the .20 or .30 cut-off, which was used as the guiding principle (Christensen et al., 2017) with only one exception where item content and difficulties were sufficiently different to allow a slightly higher partial correlation than usual. This local independence ensures that the previously reported alphas are realistic estimates of scale reliabilities as they have not been inflated by the use of redundant items.

Table 5.8. Local independence of MAP scales.

No	Scale	Avg. partial cor.	Max. partial cor.	Max. Difference
EX1	Social Need	-.14	.14	.28
EX2	Social Image	-.14	.07	.21
EX3	Work Pace	-.15	.05	.20
EX4	Risk-Taking	-.14	.10	.24
EX5	Cheerfulness	-.14	.18	.32
AG1	Trust	-.12	.11	.23
AG2	Diplomacy	-.14	.15	.28
AG3	Helpfulness	-.12	.07	.18
AG4	Compassion	-.13	.06	.18
AG5	Conflict Aversion	-.14	.14	.27
CO1	Accountability	-.15	.20	.35
CO2	Structure	-.14	.10	.23
CO3	Ambition	-.13	.16	.29
CO4	Self-Discipline	-.14	.16	.30
CO5	Decision-Making	-.14	.09	.22
ES1	Unconcern	-.16	.10	.26
ES2	Mood Stability	-.14	.11	.25
ES3	Confidence	-.08	.12	.20
ES4	Self-Control	-.15	.14	.28
ES5	Stress Tolerance	-.15	.04	.20
OP1	Imagination	-.14	.11	.25
OP2	Aesthetics	-.13	.06	.18
OP3	Self-Reflection	-.13	.11	.24
OP4	Variety	-.14	.09	.23
OP5	Mindset	-.14	-.01	.13
Mean		**-.14**	**.11**	**.24**
Median		**-.14**	**.11**	**.24**

Item invariance

In addition to the analyses of local independence, partial correlations were used to ensure that items exhibited item invariance, i.e., that there was no Differential Item Functioning (DIF) with respect to gender, age, test purpose (recruitment vs. development), or job level (managers vs. employees). Hence, response patterns of different demographic groups should be similar at comparable levels of the latent trait. Preventing DIF is a crucial aspect of ensuring a fair and unbiased assessment as well as obtaining estimates of group differences that can be attributed solely to item impact (as opposed to item bias).

Several methods to investigate DIF have been proposed, framed within Classical Test Theory (CTT) as well as Item Response Theory (IRT) (Rouquette et al., 2019; Woods, 2009). The use of partial correlations has the advantage of being easy to compute, requiring modest sample sizes for subgroups, and having a straightforward interpretation in terms of effect sizes (Stricker, 1982; Conoly, 2003).

A range of criteria to detect DIF has been suggested, none of which are universally accepted. For instance, the use of significance tests suffers the drawback that p-values become smaller as the sample size increases, in which case negligible DIF can be flagged for significance in larger samples, or substantial DIF can be overlooked in small samples. For this reason, a combination of significance and correlation magnitude was used to exclude items with substantial DIF. In addition, any demographic variable should explain less than 5 % of the variance in any item when accounting for the scale score (as proposed by Scott et al., 2010).

Table 5.9 shows the absolute average partial correlation between items and each of four demographic variables (gender, age, purpose, job level). In general, partial correlations with any demographic variable were low (< .10) averaging .06 for gender, .07 for age, .05 for purpose, and .06 for job level. In addition, the maximum partial correlation for any item with any demographic across scales was .22, which is equivalent to an explained variance of no more than 4.8 %.

Table 5.9. Item invariance of MAP scales.

No	Scale	Gender	Age	Purpose	Job level
EX1	Social Need	.05	.10	.07	.04
EX2	Social Image	.05	.05	.09	.05
EX3	Work Pace	.06	.08	.06	.04
EX4	Risk-Taking	.07	.06	.06	.04
EX5	Cheerfulness	.06	.09	.04	.07
AG1	Trust	.06	.04	.04	.04
AG2	Diplomacy	.05	.08	.06	.03
AG3	Helpfulness	.05	.07	.05	.08
AG4	Compassion	.05	.08	.06	.09
AG5	Conflict Aversion	.07	.04	.02	.06
CO1	Accountability	.08	.11	.06	.10
CO2	Structure	.08	.03	.06	.08
CO3	Ambition	.05	.09	.04	.07
CO4	Self-Discipline	.05	.08	.02	.06
CO5	Decision-Making	.07	.07	.05	.05
ES1	Unconcern	.06	.05	.07	.04
ES2	Mood Stability	.07	.07	.05	.07
ES3	Confidence	.06	.05	.04	.02
ES4	Self-Control	.05	.07	.06	.05
ES5	Stress Tolerance	.04	.05	.04	.04
OP1	Imagination	.08	.06	.03	.07
OP2	Aesthetics	.05	.04	.03	.04
OP3	Self-Reflection	.08	.09	.04	.05
OP4	Variety	.07	.08	.03	.06
OP5	Mindset	.04	.10	.04	.07
Mean		**.06**	**.07**	**.05**	**.06**
Median		**.06**	**.07**	**.05**	**.05**

Scale characteristics

Finally, a range of analyses were conducted to investigate the properties of the final scale scores. Specifically, the mean, median, standard deviation (SD), and measures of skewness and kurtosis were computed for each scale. In addition, histograms were inspected visually to ensure that scales were properly normally distributed.

The results of these analyses are listed in Table 5.10 alongside the sample size (N) used to validate the final scales. Across scales, means and medians were almost identical reflecting that scales had negligible amounts of skewness. In addition, the absolute values for both skewness and kurtosis were well below the typically used critical values of 2 (or even 1) and 4, respectively (Kim, 2013; Chissom, 1970). When using these scores to calculate normed scores (C scores), no observed proportions deviated more than 5 % from the expected percentages.

Table 5.10. Scale characteristics of MAP scales.

No	Scale	N	Mean	Median	SD	Skew.	Kurt.
EX1	Social Need	519	19.9	20	5.01	0.09	-0.51
EX2	Social Image	519	19.6	20	4.92	0.00	-0.38
EX3	Work Pace	679	18.9	19	4.40	0.25	-0.44
EX4	Risk-Taking	679	18.9	19	4.68	0.11	-0.12
EX5	Cheerfulness	807	21.7	22	4.20	-0.20	-0.39
AG1	Trust	619	20.7	21	4.99	-0.05	-0.68
AG2	Diplomacy	722	19.2	19	4.46	0.17	-0.16
AG3	Helpfulness	722	20.2	20	4.40	0.07	-0.47
AG4	Compassion	616	20.9	21	4.52	-0.15	-0.32
AG5	Conflict Aversion	619	19.4	19	4.21	0.15	-0.13
CO1	Accountability	620	19.2	19	4.32	-0.07	-0.11
CO2	Structure	522	20.2	20	4.44	-0.21	-0.44
CO3	Ambition	707	18.0	18	4.47	0.15	-0.31
CO4	Self-Discipline	620	20.9	21	4.51	-0.12	-0.38
CO5	Decision-Making	546	19.5	19	4.06	0.08	-0.16
ES1	Unconcern	504	20.9	21	4.58	-0.11	-0.42
ES2	Mood Stability	581	18.5	18	3.95	0.26	-0.01
ES3	Confidence	601	21.1	21	4.41	-0.13	-0.18
ES4	Self-Control	504	18.2	19	4.38	0.05	-0.50
ES5	Stress Tolerance	915	18.9	19	4.04	0.25	-0.26
OP1	Imagination	508	18.6	19	4.55	0.02	-0.46
OP2	Aesthetics	588	20.6	20	4.85	-0.05	-0.29
OP3	Self-Reflection	659	20.7	21	5.00	-0.24	-0.55
OP4	Variety	659	19.5	19	4.05	0.04	-0.22
OP5	Mindset	671	20.3	20	4.52	0.13	-0.26

Group differences

The validity of MAP is further supported by group differences that replicate previous findings in the research literature. Group differences were examined for gender (male/female), age (below/above 40), test purpose (development/selection), and job level (employee/manager). When constructing scales, we aimed to replicate these differences to ensure validity but keep effect sizes as low as possible to prevent scores from negatively impacting hiring rates for different demographic groups (elaborated below in the section on Adverse Impact).

Group differences on gender are displayed below in Table 5.11. Effect sizes were small (< .50) or negligible (< .20) ranging from 0.01 to 0.37 with an absolute average of 0.19. Females scored significantly higher on the Extraversion facets of Social Need and Cheerfulness, all facets in Agreeableness (with a trend towards Trust), and the Openness facet of Self-Reflection (and a trend for Aesthetics). Conversely, males scored significantly higher on the Extraversion facet of Risk-Taking, the Conscientiousness facet of Ambition, all facets in Emotional Stability (including a trend towards Unconcern), and the Openness facet of Imagination (and a trend for Mindset).

In general, these results are well in line with research showing small, yet consistent cross-cultural gender differences on the Big Five personality traits (Kajonius & Johnson, 2018).

Table 5.11. Group differences on gender (male/female) across MAP scales.

No	Scale	Male			Female			Comparison		
		N	M	SD	N	M	SD	Dif.	t	d
EX1	Social Need	301	19.4	4.80	218	20.7	5.19	-1.3	-2.96**	0.26
EX2	Social Image	301	19.7	4.76	218	19.4	5.15	0.3	0.73	0.07
EX3	Work Pace	395	18.9	4.38	284	18.9	4.43	0.0	-0.13	0.01
EX4	Risk-Taking	395	19.4	4.87	284	18.2	4.32	1.2	3.43**	0.27
EX5	Cheerfulness	438	21.1	4.14	369	22.5	4.16	-1.4	-4.81**	0.34
AG1	Trust	360	20.4	4.89	259	21.1	5.10	-0.7	-1.71	0.14
AG2	Diplomacy	396	18.5	4.40	326	20.0	4.40	-1.6	-4.73**	0.35
AG3	Helpfulness	396	19.8	4.24	326	20.6	4.56	-0.8	-2.51*	0.19
AG4	Compassion	379	20.4	4.36	237	21.7	4.65	-1.4	-3.67**	0.30
AG5	Conflict Aversion	362	18.7	4.02	257	20.3	4.29	-1.6	-4.77**	0.39
CO1	Accountability	335	19.2	4.09	285	19.1	4.57	0.1	0.17	0.01
CO2	Structure	293	20.2	4.27	229	20.2	4.66	0.02	0.05	0.01
CO3	Ambition	393	18.5	4.52	314	17.3	4.33	1.2	3.46**	0.26
CO4	Self-Discipline	335	20.7	4.52	285	21.3	4.47	-0.6	-1.66	0.13
CO5	Decision-Making	331	19.4	4.02	215	19.6	4.13	-0.2	-0.66	0.06
ES1	Unconcern	291	21.2	4.48	213	20.5	4.70	0.7	1.69	0.15
ES2	Mood Stability	328	18.9	4.17	253	18.1	3.59	0.8	2.54*	0.21
ES3	Confidence	340	21.7	4.26	261	20.4	4.50	1.3	3.71**	0.31
ES4	Self-Control	291	18.9	4.29	213	17.3	4.34	1.6	4.14**	0.37
ES5	Stress Tolerance	481	19.5	4.10	434	18.3	3.88	1.2	4.46**	0.30
OP1	Imagination	289	19.0	4.38	219	18.1	4.71	1.0	2.36*	0.21
OP2	Aesthetics	360	20.3	4.95	228	21.0	4.66	-0.7	-1.66	0.14
OP3	Self-Reflection	388	20.1	5.08	271	21.4	4.79	-1.3	-3.32**	0.26
OP4	Variety	388	19.5	4.13	271	19.6	3.93	-0.2	-0.48	0.04
OP5	Mindset	368	20.6	4.47	303	19.9	4.57	0.7	1.98	0.15

** p < .05*
*** p < .01*

Group differences for age are listed in Table 5.12. Across scales, effect sizes were small (< .50) or negligible (< .20) ranging from 0.01 to 0.48 with an average of 0.20. On average, younger individuals (below 40) scored significantly higher on the Agreeableness facets of Diplomacy, Helpfulness and Compassion, the Conscientiousness facets of Structure, Ambition and Self-Discipline, the Emotional Stability facet of Stress Tolerance, and the Openness facets of Self-Reflection and Mindset. Conversely, older individuals (40 years and above) scored significantly higher on the Extraversion facets of Social Image, the Agreeableness facet of Trust, the Emotional Stability facets of Unconcern and Confidence, and the Openness facet of Variety.

Although some of these differences contradict the maturity principle stating that levels of Agreeableness, Conscientiousness, and Emotional Stability tend to increase over the life span (Roberts et al., 2006), research using cross-sectional data (like the one presented here) on cohorts shows that younger generations tend to show higher levels of Agreeableness, Conscientiousness, and Emotional Stability compared to earlier generations (Smits et al., 2011). In addition, most research tends to focus on trait-level differences, potentially overlooking differences at the facet level (e.g., that older individuals tend to score higher on Social Image, although there is a flat trend for Extraversion in general). Another explanation for some of the contradictory findings might be contextualization, i.e., adaptations made to capture personality in a working context. For instance, younger workers might be less inclined to take risks, try new things or make somewhat risky decisions due to having less experience and lower seniority in the workplace.

Table 5.12. Group differences on age across MAP scales.

No	Scale	Below 40			Above 40			Comparison		
		N	M	SD	N	M	SD	Dif.	t	d
EX1	Social Need	274	19.9	5.04	245	19.9	4.98	0.0	-0.07	0.01
EX2	Social Image	274	19.0	4.84	245	20.3	4.93	-1.3	-3.05*	0.27
EX3	Work Pace	356	18.9	4.62	323	18.9	4.13	-0.1	-0.22	0.02
EX4	Risk-Taking	356	18.8	4.41	323	19.1	4.97	-0.3	-0.86	0.07
EX5	Cheerfulness	419	21.8	4.21	388	21.6	4.20	0.2	0.70	0.05
AG1	Trust	339	19.8	4.86	280	21.9	4.90	-2.1	-5.31**	0.43
AG2	Diplomacy	386	19.7	4.63	336	18.5	4.17	1.2	3.69**	0.28
AG3	Helpfulness	386	20.7	4.32	336	19.5	4.41	1.2	3.78**	0.28
AG4	Compassion	325	21.3	4.45	291	20.5	4.56	0.8	2.27*	0.18
AG5	Conflict Aversion	352	19.7	4.25	267	19.0	4.13	0.7	2.01*	0.16
CO1	Accountability	340	19.4	4.47	280	19.0	4.11	0.4	1.17	0.09
CO2	Structure	275	20.8	4.36	247	19.6	4.45	1.2	3.21**	0.28
CO3	Ambition	408	18.7	4.48	299	17.0	4.30	1.7	5.00**	0.38
CO4	Self-Discipline	340	21.3	4.71	280	20.5	4.22	0.8	2.10*	0.17
CO5	Decision-Making	292	19.7	3.96	254	19.2	4.16	0.5	1.55	0.13
ES1	Unconcern	278	20.1	4.50	226	21.9	4.51	-1.7	-4.33**	0.39
ES2	Mood Stability	330	18.6	3.94	251	18.5	3.96	0.1	0.24	0.02
ES3	Confidence	334	20.2	4.30	267	22.3	4.30	-2.1	-5.82**	0.48
ES4	Self-Control	278	18.5	4.36	226	17.9	4.39	0.7	1.72	0.15
ES5	Stress Tolerance	506	19.2	3.98	409	18.6	4.08	0.7	2.57*	0.17
OP1	Imagination	284	18.9	4.39	224	18.2	4.71	0.7	1.78	0.16
OP2	Aesthetics	311	20.7	4.86	277	20.4	4.84	0.3	0.82	0.07
OP3	Self-Reflection	336	21.2	4.96	323	20.2	5.00	1.0	2.55*	0.20
OP4	Variety	336	19.1	3.74	323	20.0	4.30	-0.9	-2.92**	0.23
OP5	Mindset	362	20.8	4.71	309	19.6	4.21	1.2	3.41**	0.26

* $p < .05$

** $p < .01$

Table 5.13 lists score differences between development and selection. The main purpose of these analyses was to ensure that differences were sufficiently small to warrant the use of a single norm group across different test purposes (i.e., stakes). In addition, large group differences in favor of selection might suggest that items are too socially desirable, thus making them susceptible to faking (or too undesirable when favoring development). Unsurprisingly, mean scores in selection were higher for all Emotional Stability and most Conscientiousness scales, whereas results were more mixed for scales in Extraversion, Agreeableness, and Openness. The most notable differences were higher scores in selection for Self-Discipline, Unconcern and Mindset and lower scores for Work Pace. However, most effect sizes were small (< .50) or negligible (< .20), ranging from 0.04 to 0.66 with an average of 0.30. Furthermore, there was an almost 50/50 split between differences favoring selection and development with average effect sizes of 0.33 and 0.28, respectively.

Table 5.13. Group differences on test purpose across MAP scales.

No	Scale	Development			Selection			Comparison		
		N	M	SD	N	M	SD	Dif.	t	d
EX1	Social Need	69	19.5	5.33	450	20.0	4.96	-0.5	-0.70	0.09
EX2	Social Image	69	19.9	5.36	450	19.6	4.86	0.3	0.49	0.06
EX3	Work Pace	89	21.0	4.67	590	18.6	4.27	2.4	4.81**	0.55
EX4	Risk-Taking	89	19.3	5.62	590	18.9	4.53	0.4	0.76	0.09
EX5	Cheerfulness	54	20.3	5.00	753	21.8	4.13	-1.5	-2.60**	0.37
AG1	Trust	82	20.0	4.94	537	20.8	4.99	-0.9	-1.47	0.17
AG2	Diplomacy	52	20.2	5.59	670	19.1	4.36	1.1	1.65	0.24
AG3	Helpfulness	52	21.9	5.11	670	20.0	4.32	1.9	3.02**	0.44
AG4	Compassion	76	21.3	5.47	540	20.8	4.37	0.5	0.84	0.10
AG5	Conflict Aversion	60	19.2	4.60	559	19.4	4.17	-0.2	-0.30	0.04
CO1	Accountability	47	20.6	4.78	573	19.1	4.26	1.6	2.40*	0.36
CO2	Structure	74	18.8	4.56	448	20.4	4.39	-1.6	-2.91**	0.37
CO3	Ambition	59	18.2	4.64	648	18.0	4.46	0.2	0.33	0.04
CO4	Self-Discipline	47	18.3	5.11	573	21.2	4.39	-2.9	-4.27**	0.65
CO5	Decision-Making	54	18.3	4.75	492	19.6	3.96	-1.3	-2.29*	0.33
ES1	Unconcern	21	18.0	4.76	483	21.0	4.54	-3.0	-2.94**	0.66
ES2	Mood Stability	64	17.3	3.98	517	18.7	3.91	-1.5	-2.79**	0.37
ES3	Confidence	47	20.2	4.92	554	21.2	4.36	-1.0	-1.46	0.22
ES4	Self-Control	21	16.9	5.32	483	18.3	4.33	-1.4	-1.43	0.32
ES5	Stress Tolerance	51	18.8	3.44	864	19.0	4.07	-0.2	-0.29	0.04
OP1	Imagination	22	20.5	5.24	486	18.5	4.50	2.0	2.04*	0.44
OP2	Aesthetics	32	21.9	5.08	556	20.5	4.83	1.4	1.54	0.28
OP3	Self-Reflection	57	22.3	5.49	602	20.5	4.93	1.8	2.63**	0.36
OP4	Variety	57	20.9	4.73	602	19.4	3.96	1.5	2.70**	0.37
OP5	Mindset	33	17.6	3.71	638	20.4	4.52	-2.8	-3.54**	0.63

* *p < .05*
** *p < .01*

Finally, we inspected differences in mean scores between employees (including specialists) and managers (including both middle managers and executives) listed below in Table 5.14. Most notably, managers scored markedly higher than employees on Social Image and lower on Diplomacy as well as Conflict Aversion, which should come as no surprise given the usual requirements for management positions (i.e., being visible, direct, decisive, giving honest feedback, etc.). At the trait level, although some mixed patterns emerged, managers tended to score higher on Extraversion, Emotional Stability and Openness, whereas employees scored higher on Agreeableness and Conscientiousness (with Conscientiousness and Openness showing most mixed results at the facet level). In general, these results replicate research findings on the relation between personality traits and different management levels (Kang et al., 2023; Cuppello et al., 2023).

Table 5.14. Group differences on job level across MAP scales.

No	Scale	Employee			Manager			Comparison		
		N	M	SD	N	M	SD	Dif.	t	d
EX1	Social Need	387	19.7	5.12	113	20.5	4.75	-0.8	-1.41	0.19
EX2	Social Image	387	18.8	5.06	113	21.9	3.70	-3.1	-6.06**	0.79
EX3	Work Pace	532	18.7	4.36	126	19.7	4.39	-1.1	-2.48*	0.27
EX4	Risk-Taking	532	18.5	4.70	126	20.3	4.30	-1.8	-3.85**	0.45
EX5	Cheerfulness	555	21.6	4.27	134	22.3	3.99	-0.7	-1.66	0.16
AG1	Trust	465	20.6	5.02	135	20.7	5.02	-0.1	-0.21	0.03
AG2	Diplomacy	545	19.8	4.40	138	16.8	4.05	2.9	7.04**	0.67
AG3	Helpfulness	545	20.5	4.45	138	18.8	3.82	1.7	4.21**	0.44
AG4	Compassion	478	20.9	4.51	104	20.6	4.58	0.3	0.66	0.07
AG5	Conflict Aversion	513	19.8	4.24	99	17.2	3.29	2.6	5.69**	0.62
CO1	Accountability	495	19.1	4.37	109	19.4	3.86	-0.3	-0.59	0.06
CO2	Structure	332	20.7	4.27	90	18.9	4.58	1.8	3.48**	0.41
CO3	Ambition	451	18.1	4.52	105	18.4	4.61	-0.3	-0.70	0.08
CO4	Self-Discipline	495	21.2	4.49	109	19.5	4.47	1.7	3.55**	0.38
CO5	Decision-Making	394	19.8	4.10	112	18.2	3.68	1.6	3.65**	0.39
ES1	Unconcern	353	20.6	4.46	89	22.6	4.61	-2.0	-3.78**	0.45
ES2	Mood Stability	395	18.5	3.89	73	17.9	3.75	0.6	1.23	0.16
ES3	Confidence	470	20.7	4.29	99	22.6	4.71	-1.9	-3.63**	0.48
ES4	Self-Control	353	18.4	4.48	89	17.8	3.93	0.6	1.18	0.14
ES5	Stress Tolerance	639	18.9	4.06	111	19.9	4.00	-1.0	-2.35*	0.24
OP1	Imagination	385	18.4	4.57	103	19.2	4.39	-0.8	-1.55	0.17
OP2	Aesthetics	451	20.7	4.79	108	20.2	5.20	0.6	1.10	0.12
OP3	Self-Reflection	485	20.7	4.99	150	20.6	5.11	0.1	0.31	0.03
OP4	Variety	485	19.1	3.95	150	20.7	4.04	-1.5	-4.17**	0.39
OP5	Mindset	519	20.5	4.62	133	19.7	4.17	0.8	1.78	0.17

* $p < .05$

** $p < .01$

Criterion validity

Criterion validity refers to the relationship between test results and information about test subjects derived from other sources (i.e., external criteria).

Concurrent validity: Managerial performance

Following the initial development of MAP, a concurrent validity study was conducted linking test scores to a measure of performance in a group of Swedish managers. The validation sample consisted of 73 managers from two administrations of a large Swedish municipality. The mean age in the sample was 44 years (SD = 9) and 58% were female. The level of education ranged from elementary school to postgraduate education, with a majority (57%) having completed at least three years of high school education and/or some form of tertiary education.

In the validation study, two summarized indexes constituted the overall criterion Managerial Performance. A rating questionnaire with 28 items was designed to describe the employee's (in this case the manager) behaviors at work. The questionnaire included items such as "the manager/ leader carries out the work carefully and thoroughly" and "the manager/leader has a positive attitude" rated on a 4-point Likert scale (1 = Strongly disagree, 2 = Partially disagree, 3 = Partially agree, 4 = Strongly agree). The ratings were carried out by each manager's supervisor. In a factor analysis including all 28 items, a strong factor emerged with an Eigenvalue value of 12, explaining 43% of the total variance. Based on these results, it was decided to summarize all items into an overall index representing the criterion of Supervisory Performance Rating of managerial performance. The reliability (Cronbach's alpha) of this overall Supervisory Performance Rating was estimated to.94.

The managers' performance was also measured using subordinates' ratings. Data underlying this measure was collected in an annual employee survey at the municipality. Items regarding their managers' behaviors were administered to all employees. Item examples include:

- My closest manager is good at planning and organizing work
- I know what my closest manager expects of me with regards to my work
- My immediate boss makes sure that what we decided to do really gets done
- My closest manager communicates openly and honestly.

The questions were answered on a 5-point Likert scale where 1 corresponded to "Strongly disagree" and 5 corresponded to "Strongly agree". A factor analysis including all 8 items showed one strong factor with an Eigenvalue of 5, explaining 67% of the total variance. Based on these results, it was decided to summarize all items into an overall index representing the criteria of Subordinates Performance Rating. The reliability (Cronbach's alpha) of this Subordinates Performance Rating was estimated to .94.

Finally, the Supervisory and Subordinates Performance Ratings were summarized into an overall index representing the overall criterion of Managerial Performance (Cronbach's alpha = .94). An assumption was made that the two indexes partly measure different constructs: different questions were asked to the subordinate and the supervisor and there were different assessors. As expected, the correlation between the two indexes was not significantly different from zero (r =.12, p < .05), implying the two measures were supplementary to each other.
Validity studies of this type may underestimate the validity due to unreliability of the criterion (performance ratings) and hence need to be corrected. As employees were rated by only a single manager, it was not possible to estimate inter-rater reliability. Therefore, an estimate of .52 from a recognized and comprehensive meta-analysis on performance ratings was applied (Viswesvaran, Ones & Schmidt, 1996).

Besides criterion unreliability, the validity estimate might also be underestimated as organizations rarely select individuals for employment purely by chance but based on who is most likely to perform. Therefore, it is highly likely that there is limited variation in the data, thus attenuating the validity estimate (known as indirect range restriction). To determine the level of range restriction (U), the variation in the standardization sample of the overall managerial performance composite score (SD = 0.47) was compared to that of the 73 managers who participated in the study (SD = 0.35), yielding an estimated range restriction of U = 0.74.

Table 5.15 shows the observed correlation (r) between the overall Managerial Performance Composite Score (MPCS) and the overall criterion of Managerial Performance and the estimate of operational validity corrected for criterion unreliability and range restriction alongside the 95 % confidence interval based on the calculations presented by Hunter & Schmidt (2004). The Managerial Performance Composite Score (MPCS) was obtained by mechanically weighting each of the trait scores into a composite score by means of a multiple linear regression, the results of which are shown in the first part of Table 5.11. In conclusion, the study shows that when scales are combined optimally, MAP is a strong predictor of managerial performance as rated by both supervisors and subordinates.

Table 5.15. Operational validity of trait-based Managerial Performance Composite Score (MCPS)

MAP trait	Std. Beta		
Extraversion	.14		
Agreeableness	-.01		
Conscientiousness	.19		
Emotional Stability	-.04		
Openness	-.02		
Operational validity	r	p	$CI_{.95}$
MCPS	.31	.54	.16-.86

Concurrent validity: Norwegian retail workers

In 2012, a concurrent criterion-related validity study was conducted using the Norwegian version of MAP. This sample consisted of employees within the grocery store sector in Norway. In total, 36 store managers assessed 130 employees. Each store manager rated the employee's overall work performance on 9 questions (e.g., "solves problems on their own ","is competitive" and "remembers important things") using a Likert-scale.

When evaluating the properties of the job performance ratings, it became evident that a bias existed between the raters (store managers). It was considered likely that the bias was due to systematic measurement errors as store managers may only compare employees working at a single workplace (store). It was therefore decided to only use ratings from store managers who had rated more than two employees. Further, the ratings used were standardized within each rater. In the analysis, each store manager thus represents a unit and is treated as a separate sample. The scores from each unit were used as the criterion in the analysis.

Validity studies of this type may underestimate the validity due to unreliability of the criterion (performance ratings) and hence need to be corrected. As employees were rated by only a single manager, it was not possible to estimate inter-rater reliability. Therefore, an estimate of .52 from a recognized and comprehensive meta-analysis on performance ratings was applied (Viswesvaran, Ones & Schmidt, 1996). Besides criterion unreliability, validity might also be underestimated as organizations rarely select individuals for employment purely by chance but based on who is most likely to perform. Therefore, it is highly likely that there is limited variation in the data, thus attenuating the validity estimate. This phenomenon is known as (indirect) range restriction but can be estimated and corrected for when conducting these types of validity studies.

To investigate the restriction of range, the standard deviation (SD) for the five traits in the standardization sample (2) was compared to the variation in the validation sample. By dividing the SD from the validation sample with the SD from the standardization sample, one can estimate the range restriction (U) of each trait. The estimates may then be applied to correct the validity (Schmidt, Shaffer, & Oh, 2008).

The results of the study are presented below in Table 5.16 supporting the criterion validity of MAP regarding job performance. As expected, all traits correlated positively with job performance, with Conscientiousness showing the highest correlation followed by Extraversion and Emotional Stability.

Table 5.16. Operational validity of MAP for Norwegian retail workers (N = 130).

MAP trait	U	r	p
Extraversion	0.91	.14	.22
Agreeableness	0,92	.02	.04
Conscientiousness	0.89	.15	.23
Emotional Stability	0.89	.12	.19
Openness	0.85	.05	.09

Note. U = Restriction of range, r = observed correlation, p = Operational validity corrected for criterion unreliability and range restriction.

Concurrent validity: Swedish insurance employees

In 2019, a concurrent criterion-related validity study was conducted using the Swedish version of MAP in collaboration with an insurance company in Sweden. Job performance was assessed by either the employee's manager or a person in a similar position within the company. The employees were evaluated on their overall work performance based on two criteria: Future potential for the organization and daily contribution to the organization. The managers rated the employee's performance using several standardized statements, each with three different rating levels – low, medium and high. The data collected using the assessments of potential and contribution were combined into a single overall estimate that was analyzed together with the employee's MAP result.

In total, 87 employees were assessed over a six-month period. The sample comprised 52 % males and 48 % females aged 18-65 (with most employees in the age category 18-30 followed by 31-40). Participants reported either high school (44 %), less than three years of post-secondary education (22 %) or more than three or more years of post-secondary education (25 %) as their highest completed education (for the remaining 9 %, "other education" was stated, or no information was given.

Table 5.17 shows the observed and corrected correlations between the combined performance rating and the five personality traits. The results provide support for the criterion validity of MAP regarding job performance as all factors correlated positively with the performance rating, with Conscientiousness showing the strongest correlation.

Table 5.17. Operational validity of MAP for Swedish insurance employees (N = 87).

MAP trait	U	r	p
Extraversion	1.02	.28	.27
Agreeableness	1.02	.07	.07
Conscientiousness	0.90	.32	.36
Emotional Stability	0.87	.16	.18
Openness	0.85	.14	.17

Note. U = Restriction of range, r = observed correlation, p = Operational validity corrected for criterion unreliability and range restriction.

Concurrent validity: Performance dimensions in a sample of German workers
Finally, in 2023, a study was carried out in collaboration with the Justus-Liebig University to predict different performance dimensions in a sample of 254 workers in Germany who completed MAP and a performance questionnaire through the online Prolific platform (please note that this study was also used to cross-validate MAP-X).

The sample comprised 44.5 % male and 55.1 % female workers aged 20-74 (M = 42.0, SD = 12.1). Of the 254 participants, 51.6 % were employees and 48.4 % were managers working in a diverse range of work areas (administration, customer service, healthcare, finance, sales etc.) and with different educational backgrounds (with the majority of 35.0 % having a bachelor's degree or equivalent as their highest level of education).

Besides MAP, participants completed a set of questionnaires assessing different performance dimensions:

- Counterproductive Work Behavior (CWB): CWB was measured with 10-item short version of the Counterproductive Work Behavior Checklist (CWB-C) assessing a range of different counterproductive work behaviors such as coming in late without permission or starting arguments with coworkers (Spector et al., 2010).
- Organizational Citizenship Behavior (OCB): OCB was assessed with a 25-item questionnaire (Staufenbiel & Hartz, 2000) assessing the facets of altruism (helping others), conscientiousness (following rules and instructions), sportsmanship (handling changes and unforeseen events without complaining), and civic virtue (attending conferences, improving skills, representing the company well outside of daily work, etc.).
- Task Performance (TP): The OCB questionnaire also contained a subset of items assessing in-role (task) performance, i.e., completing tasks and meeting formal requirements of the job.

The correlations between MAP traits and performance dimensions are listed below in Table 5.18 with the highest correlation for each dimension highlighted in bold. The last row in the table shows the multiple correlation coefficient (R) derived from a series of multiple linear regressions predicting performance dimensions using optimal weighting (i.e., linear combinations) of trait scores.

In line with expectations, Conscientiousness showed the highest correlations with Task Performance, the OCB facet of conscientiousness and total performance (excluding CWB). Agreeableness was most predictive of CWB as well as total OCB, especially the facet of altruism. The final OCB facets of sportsmanship and civic virtue were best predicted by Emotional Stability and Extraversion, respectively.

Surprisingly, CWB correlated higher with Agreeableness than either Conscientiousness or Emotional Stability as would be expected based on prior research. However, the CWB scale used in the current study had many items relating to interpersonal aspects such as arguing with, insulting or making fun of others (as opposed to other questionnaires measuring theft, absenteeism, unsafe behavior, etc.).

Table 5.18. Correlations between MAP traits and performance dimensions in a sample of German workers (N = 254).

MAP trait	CWB	ALT	CON	SPO	CIV	OCB	TP	TOTAL
Extraversion	-.01	.37	.08	.11	**.47**	.38	.08	.35
Agreeableness	**-.36**	**.50**	.31	.34	.38	**.55**	.31	.54
Conscientiousness	-.25	.31	**.46**	.25	.43	.53	**.46**	**.56**
Emotional Stability	-.33	.21	.33	**.42**	.28	.45	.31	.46
Openness	-.01	.20	-.11	.08	.34	.20	.02	.17
Multiple R	**.49**	**.58**	**.53**	**.51**	**.62**	**.68**	**.53**	**.69**

Note. CWB = Counterproductive Work Behavior, ALT = Altruism, CON = Conscientiousness,
SPO = Sportsmanship, CIV = Civic Virtue, OCB = Organizational Citizenship Behavior, TP = Task Performance.

In conclusion, these studies support the criterion-related validity of MAP across a range of performance dimensions and occupations.

6. Reliability

Reliability is defined as the consistency with which an instrument measures a construct.

An often-used measure of internal consistency is Cronbach's alpha (Cronbach, 1951), which is listed below for each of the MAP scales in Table 6.1 based on the sample representing the norm group.

The final column of the table contains the Standard Error of Measurement (SEM) defined as:

$$SEM = SD * \sqrt{(1-r)}$$

Where SD represents the standard deviation, and r refers to the reliability of the scale in question. As shown in Table 13, all scales have acceptable or excellent levels of reliability with alphas ranging from .70 to .86 and an average of .76. The Standard Error of Measurement (for the raw score) ranged from 1.75 to 2.39 with an average of 2.19. When used to construct a 95 % confidence interval, the true score (T) would most likely fall within a range of no more than ±4.3 points from the observed score (O). When normed and converted to C-scores (with an SD of 2), the average SEM is only 0.98, i.e., less than one C score point.

In sum, these estimates show great consistency and measurement accuracy for the different MAP scales.

Table 6.1. Reliability and Standard Error of Measurement (SEM) of MAP scales.

No	Scale	Items	Alpha	SD	SEM	SEM (C score)
EX1	Social Need	8	.79	5.01	2.30	0.92
EX2	Social Image	8	.80	4.92	2.20	0.89
EX3	Work Pace	8	.76	4.40	2.16	0.98
EX4	Risk-Taking	8	.86	4.68	1.75	0.75
EX5	Cheerfulness	8	.74	4.20	2.14	1.02
AG1	Trust	8	.80	4.99	2.23	0.89
AG2	Diplomacy	8	.77	4.46	2.14	0.96
AG3	Helpfulness	8	.74	4.40	2.24	1.02
AG4	Compassion	8	.78	4.52	2.12	0.94
AG5	Conflict Aversion	8	.70	4.21	2.31	1.10
CO1	Accountability	8	.70	4.32	2.37	1.10
CO2	Structure	8	.71	4.44	2.39	1.08
CO3	Ambition	8	.74	4.47	2.28	1.02
CO4	Self-Discipline	8	.75	4.51	2.26	1.00
CO5	Decision-Making	8	.74	4.06	2.07	1.02
ES1	Unconcern	8	.76	4.58	2.24	0.98
ES2	Mood Stability	8	.72	3.95	2.09	1.06
ES3	Confidence	8	.74	4.41	2.25	1.02
ES4	Self-Control	8	.75	4.38	2.19	1.00
ES5	Stress Tolerance	8	.70	4.04	2.20	1.09
OP1	Imagination	8	.80	4.55	2.03	0.89
OP2	Aesthetics	8	.79	4.85	2.22	0.92
OP3	Self-Reflection	8	.81	5.00	2.18	0.87
OP4	Variety	8	.73	4.05	2.10	1.04
OP5	Mindset	8	.74	4.52	2.30	1.02
Mean		**8**	**.76**	**4.48**	**2.19**	**0.98**
Median		**8**	**.75**	**4.46**	**2.20**	**1.00**

7. Standardization

Standardization refers to the procedure of design and testing that leads to a standardized test. Standardization thus says something about the way in which the test is constructed, thoroughly tried, and tested. There are several ways to standardize, where the best known are the normative and ipsative methods. MAP is a normative test, which means that the test result is compared to a relevant norm group.

One of the advantages of normative tests is that they are quick and straightforward to complete. Although there are typically more questions compared to other types of tests, normative tests still take a short time to complete because the questions are easier to answer. Another strength of the normative method is that the test scales are completely independent of each other. Because the scales are measured one at a time, normative tests show more nuances and make the measurements more accurate.

Most importantly, normative tests are suitable for comparing individuals. Normative tests not only provide answers to what is characteristic of individuals but also what is characteristic of test persons in relation to others. Normative tests thus measure interpersonal differences (differences between people), where the person’s response is compared to the responses of others. Therefore, the normative approach is the preferred method when a tool is to be used for selection purposes and is also an ideal tool for development purposes (providing insight as to how the individual differs from others).

Score calculation

First, the responses to all items (answered on a 4-point Likert scale with the options Disagree/Slightly disagree/Slightly agree/Agree) in each scale are summed to a raw score, which is then compared to the answers from a norm group. These raw scores are then converted to z-scores by subtracting the raw score from the mean and dividing by the standard deviation of scores in the norm group. Then, z-scores are converted to C scores with a mean of 5 and a standard deviation of 2, which is displayed as the results. The interpretations, z-score ranges, percentages, and percentiles for C scores are shown below in Table 7.1.

Table 7.1. Interpretation of C scores.

Category	C score	z-score	Percentage	Percentile
Low	0	-2.75;2.25	1	1
	1	-2.25;-1.75	3	4
	2	-1.75;-1.25	7	11
	3	-1.25;-0,75	12	23
Moderate	4	-0.75;-0-25	17	40
	5	-0.25;0.25	20	60
	6	0.25;0.75	17	77
High	7	0.75;1.25	12	89
	8	1.25;1.75	7	96
	9	1.75;2.25	3	99
	10	2.25-2.75	1	100

Norm group

At Assessio, we are committed to offering norms of the highest quality based on quality standards derived from various international standards, including EFPA, COTAN, and ITC guidelines. In short, these guidelines set out criteria for various aspects of the norm group:

- Update: When was the norm group last updated?
- Sample size: How large is the norm group? Is it sufficiently large to ensure representativeness?
- Composition: How is the norm group composed with respect to different demographics?
- Subgroup differences: Are group differences sufficiently small to prevent adverse impact?

Update

Over time, what is considered normal behavior changes. Major events and crises have an impact on the way people in general behave and new generations may also challenge the existing standards. Therefore, with respect to assessments, it is highly important to update norm groups at a regular basis and make sure that all candidates and people assessed are evaluated with a norm group representing the current state and what is currently considered normal behavior, since that will provide the most valid assessment. In addition, updating the norm group keeps scores balanced and avoids too many candidates getting either high or low scores. In other words, norm updates allow for better differentiation of candidates, which in turn leads to better recruitment decisions.

According to EFPA and COTAN guidelines, a norm of the highest quality should not be older than 10 or 15 years, respectively. At Assessio, however, we are committed to checking if updates are needed at least every 2 years and update our norm groups frequently.
As the current manual reflects a revision of the current version of MAP, the initial norm group is a research-based norm group based on data collected in a high-stake setting (selection and development) from February 2024 to March 2025. Once more data is collected, the norm group will be updated accordingly.

Sample size

A good norm group consists of many people, as a high number provides greater representation and statistical certainty. The prevailing view is that the larger the sample, the better the norm group. While that is true, it very much depends on sampling procedures as well as composition with respect to different demographic characteristics. In general, norm groups that are too small run the risk of underrepresentation (e.g., too few people with a certain occupation or education level), whereas too large norm groups risk overrepresentation (e.g., too many people of a certain age or nationality). According to EFPA, a sample size of at least 1,000 constitutes an excellent norm group (in some cases, smaller norm groups may also be sufficient depending on composition, target groups, and intended applications). For high-stake purposes, a norm group consisting of 400-999 people is considered a good sample size (EFPA, 2013).

As data for the revised MAP scales were collected in succession, the sample size varies slightly between different scales as shown below in Table 147.2 ranging from 504 to 915 with an average sample size of 628. For all scales, this represents a good sample size, even in high-stake decisions, according to both EFPA and COTAN guidelines.

Composition

To ensure that a norm group is representative of all target groups and is appropriate for all intended applications, key demographic characteristics must be carefully weighed and balanced, especially those that can lead to potential score differences between subgroups.

The current research-based norm group for MAP consists of people aged 18-70 who completed the assessment in a high-stake setting (selection and development) and in their native language. As statistical analyses showed mostly negligible or small group differences for gender, age, test purpose, and job level, the norm group was not further stratified for any of these demographic variables, as this would only reduce the sample size without impacting overall scores across groups. Importantly, the majority of cases in the norm group represent the test purpose of selection (91 % on average) and the job level of employee (74 % on average) with an almost 50/50 split between males and females as well as people above and below the age of 40.

The demographic composition of the norm group for each scale is listed below in Table 7.2.

Table 7.2. Demographic composition of the norm group for MAP.

No	Scale	N	Male %	Below 40 %	Selection %	Employee %
EX1	Social Need	519	58.0	52.8	86.7	74.6
EX2	Social Image	519	58.0	52.8	86.7	74.6
EX3	Work Pace	679	58.2	52.4	86.9	78.4
EX4	Risk-Taking	679	58.2	52.4	86.9	78.4
EX5	Cheerfulness	807	54.3	51.9	93.3	68.8
AG1	Trust	619	58.2	54.8	86.8	75.1
AG2	Diplomacy	722	54.8	53.5	92.8	75.5
AG3	Helpfulness	722	54.8	53.5	92.8	75.5
AG4	Compassion	616	61.5	52.8	87.7	77.6
AG5	Conflict Aversion	619	58.5	56.9	90.3	82.9
CO1	Accountability	620	54.0	54.8	92.4	79.8
CO2	Structure	522	56.1	52.7	85.8	63.6
CO3	Ambition	707	55.6	57.7	91.7	63.8
CO4	Self-Discipline	620	54.0	54.8	92.4	79.8
CO5	Decision-Making	546	60.6	53.5	90.1	72.2
ES1	Unconcern	504	57.7	55.2	95.8	70.0
ES2	Mood Stability	581	56.5	56.8	89.0	68.0
ES3	Confidence	601	56.6	55.6	92.2	78.2
ES4	Self-Control	504	57.7	55.2	95.8	70.0
ES5	Stress Tolerance	915	52.6	55.3	94.4	69.8
OP1	Imagination	508	56.9	55.9	95.7	75.8
OP2	Aesthetics	588	61.2	52.9	94.6	76.7
OP3	Self-Reflection	659	58.9	51.0	91.4	73.6
OP4	Variety	659	58.9	51.0	91.4	73.6
OP5	Mindset	671	54.8	53.9	95.1	77.3
Mean		**628**	**57.1**	**54.0**	**91.1**	**74.1**

Group differences & Adverse Impact

When using an assessment to make important decisions with a great impact on individuals (such as selection, promotion, and hiring decisions), a key requirement is to ensure fairness and mitigate Adverse Impact (AI), defined as "a substantially different rate of selection in hiring, promotion, or other employment decisions which works to the disadvantage of members of a race, sex or ethnic group" (Uniform Guidelines on Employee Selection Procedures, Equal Employment Opportunity Commission, 1978). The "Four-Fifths rule" can be used to determine whether an assessment has AI. Usually, a selection rate for any demographic group less than four-fifths (or 80 percent) of the selection rate for the group with the highest rate (majority group) is considered evidence of AI. The level of AI depends both on the magnitude of group differences (e.g., between males and females) and the selection ratio, i.e., the number of people hired compared to the total number of applicants.

The first step in preventing Adverse Impact is to ensure that differences between demographic groups reflect true differences (item impact) and not item bias (or Differential Item Functioning, DIF). That is, including items with a uniform bias for one demographic group would inflate group differences, thus creating a disadvantage for the lowest scoring group. Ensuring that items are not subject to item bias ensure a report of true group differences. However, true group differences can still produce Adverse Impact making it important to study further.

Second, simulations of expected AI are conducted at different selection rates for gender (males/females) and age (above/below 40). Please note, however, that these calculations are based on the assumptions that 1) candidates are selected based on a single score only, 2) the assessment is used as the sole basis for selection and 3) a fixed selection rate is applied (i.e., hiring everyone scoring above a predefined cut-off). In practice, Assessio recommends basing recruitment decisions on a combination of assessments, scales, and other information relevant to the job in question (i.e., KSAOs) to consider both job, team, and organization fit.

Tables 7.3 and 7.4 list the standardized mean difference (Cohen's d) between groups alongside the simulated AI ratio (selection rate of the least represented group compared to the most represented group) for gender (male/female), and age (above/below 40), respectively. The calculations are based on three fixed selection ratios (SR): Strict (C score 7-10), Moderate (C score 6-10) and Lenient (C score 5-10) equivalent to the top 23, 40, and 60 %, respectively. For any given scale, we aimed for an AI ratio above 0.80 for a lenient selection ratio as suggested by the Four-Fifths rule.

The results show that when applying strict selection ratios, the AI ratio for many scales is below the .80 cut-off suggested by the four-fifths rule and so should be avoided, especially for AG scales (favoring females and younger individuals), CO scales (favoring younger individuals, and ES scales (favoring males and to some extent older individuals).

For lenient selection rates, the selection rate of the least represented group is no less than 80 % of the selection rate of the most represented group for most scales across demographic groups. For gender, the only exception was Conflict Aversion and Self-Control with an AI ratio of 0.77 favoring females and males, respectively. Regarding age, only one AI ratio favored younger individuals (0.74 for Ambition) and two AI ratios favored older individuals (0.75 for Unconcern and 0.70 for Confidence, respectively. When using these scales for selection, careful consideration should be given to avoid fixed selection ratios (even if lenient) and combine scores on this scale with other criteria to balance out the level of Adverse Impact, thus preventing any indirect discrimination on gender or age.

In conclusion, when applying proper selections ratios and decision rules (i.e., combining (multiple) scores with information derived from other sources), MAP provides a fair and unbiased assessment that does not cause any Adverse Impact for protected groups when used for making employment decisions.

Table 7.3. Adverse Impact simulations for gender (male/female) for different selection ratios (SR) across MAP scales.

No	Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
EX1	Social Need	-1.3	0.26	0.69	0.75	0.83
EX2	Social Image	0.3	0.07	0.71	0.81	0.91
EX3	Work Pace	0.0	0.01	0.99	1.00	0.95
EX4	Risk-Taking	1.2	0.27	0.59	0.74	0.89
EX5	Cheerfulness	-1.4	0.34	0.64	0.72	0.84
AG1	Trust	-0.7	0.14	0.74	0.93	0.87
AG2	Diplomacy	-1.6	0.35	0.67	0.66	0.80
AG3	Helpfulness	-0.8	0.19	0.78	0.78	0.91
AG4	Compassion	-1.4	0.30	0.59	0.70	0.87
AG5	Conflict Aversion	-1.6	0.39	0.46	0.65	0.77
CO1	Accountability	0.1	0.01	0.89	0.98	0.91
CO2	Structure	0.02	0.01	0.83	0.95	1.00
CO3	Ambition	1.2	0.26	0.74	0.81	0.80
CO4	Self-Discipline	-0.6	0.13	0.86	0.99	0.87
CO5	Decision-Making	-0.2	0.06	1.00	0.94	0.90
ES1	Unconcern	0.7	0.15	0.96	0.90	0.87
ES2	Mood Stability	0.8	0.21	0.75	0.78	0.82
ES3	Confidence	1.3	0.31	0.76	0.72	0.83
ES4	Self-Control	1.6	0.37	0.63	0.74	0.77
ES5	Stress Tolerance	1.2	0.30	0.68	0.75	0.81
OP1	Imagination	1.0	0.21	0.87	0.73	0.80
OP2	Aesthetics	-0.7	0.14	0.85	0.87	0.88
OP3	Self-Reflection	-1.3	0.26	0.80	0.81	0.82
OP4	Variety	-0.2	0.04	0.97	0.94	0.95
OP5	Mindset	0.7	0.15	0.91	0.86	0.85

Table7.4. Adverse Impact simulations for age (below/above 40) for different selection ratios (SR) across MAP scales.

No	Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
EX1	Social Need	0.0	0.01	0.94	0.92	0.94
EX2	Social Image	-1.3	0.27	0.94	0.89	0.89
EX3	Work Pace	-0.1	0.02	0.87	0.91	0.97
EX4	Risk-Taking	-0.3	0.07	0.72	0.79	0.95
EX5	Cheerfulness	0.2	0.05	0.97	0.94	1.00
AG1	Trust	-2.1	0.43	0.55	0.65	0.81
AG2	Diplomacy	1.2	0.28	0.71	0.76	0.82
AG3	Helpfulness	1.2	0.28	0.80	0.72	0.84
AG4	Compassion	0.8	0.18	0.85	0.94	0.90
AG5	Conflict Aversion	0.7	0.16	0.73	0.74	0.84
CO1	Accountability	0.4	0.09	0.82	0.84	0.96
CO2	Structure	1.2	0.28	0.74	0.80	0.81
CO3	Ambition	1.7	0.38	0.46	0.64	0.74
CO4	Self-Discipline	0.8	0.17	0.64	0.83	0.94
CO5	Decision-Making	0.5	0.13	0.81	0.81	0.96
ES1	Unconcern	-1.7	0.39	0.67	0.64	0.75
ES2	Mood Stability	0.1	0.02	0.94	0.99	0.97
ES3	Confidence	-2.1	0.48	0.55	0.60	0.70
ES4	Self-Control	0.7	0.15	0.80	0.83	0.84
ES5	Stress Tolerance	0.7	0.17	0.92	0.82	0.85
OP1	Imagination	0.7	0.16	0.95	0.91	0.85
OP2	Aesthetics	0.3	0.07	0.83	0.94	0.88
OP3	Self-Reflection	1.0	0.20	0.76	0.81	0.88
OP4	Variety	-0.9	0.23	0.62	0.80	0.90
OP5	Mindset	1.2	0.26	0.65	0.77	0.80

8. Translations and Adaptations

The following section outlines the process for translations and adaptation of items. The full list of available languages is available at the Assessio help center (https://helpcenter.assessio.com/hc/en-us/articles/18233719887517-Languages).

General procedure for translations and adaptations

When translating an existing instrument from the original (source) language into a new (target) language, it is crucial to consider two different aspects: Linguistic similarity and psychological similarity. Linguistic (or semantic) similarity is often established using literal translations with the purpose of developing a test as close to the original as possible with respect to language characteristics such as wording, phrases, and connotations. In the literature, this method is variously termed "translation" or "application" (Van de Vijver & Poortinga, 2005).

However, enhancing linguistic similarity sometimes occurs at the expense of psychological similarity, defined as individual test items having the same meaning, interpretation, and level of difficulty across languages. The process of changing certain items to ensure psychological similarity is commonly referred to as "adaptation" (Van de Vijver & Poortinga, 2005). These changes span the range of minor adjustments (such as changing the currency on items of numerical skills) to entire revisions, i.e. rewriting items completely to capture the intended meaning (as opposed to opting for a literal, yet non-idiomatic translation).

Translational protocol

When translating/adapting our assessment tools, we aim at enhancing both linguistic and psychological similarity. Wherever possible, literal translations of items are preserved with only minor adjustments made when needed. When literal translations do not suffice, alterations are made to adapt to the specific target language and culture. In this regard, our procedures closely follow expert recommendations on conducting adaptions and are aligned with the criteria of EFPA's test review model (Van de Vijver & Poortinga, 2005; EFPA, 2013).

Our translations/adaptations follow a strict protocol with the following steps:

1) **Selection of translators and reviewers**
   Translators and reviewers are certified native language translators residing in the country of the target language, which ensures high quality, authorized translations.
2) **Briefing of translators**
   Translators are given thorough instructions on translational procedures and principles with respect to wording, style, and localization (e.g., selecting proper currencies, metrics, names, idiomatic language, etc.). The full list of instructions is outlined below.
3) **First translation**
   The first translation is always based on the English source version (or multiple source versions, if possible).
4) **First proof reading**
   A proof reading of the first translation is made by a different certified translator.
5) **Back-translation**
   The most recent version is back-translated into the source language aimed specifically at ensuring linguistic similarity and identifying necessary adaptations (i.e., discrepancies between the target and source language versions).
6) **First revision**
   Any discrepancies and adaptions made to the target language versions are reviewed by our own team of subject matter experts. Then, adjustments are made in close collaboration with

the initial translator to ensure equivalence with the source language version and that ordinary, idiomatic language is used in the target language version.

7) **First implementation**
The first version is implemented in the test platform.
8) **Functional test**
From the test platform, a certified native-speaking translator makes a proof reading and suggests adjustments in relation to the target context.
9) **Second revision**
Based on the functional test, the translator suggests adjustments that are discussed with our team of subject matter experts to arrive at a second version.
10) **Second implementation**
The second version is implemented in the test platform and made available to candidates.
11) **Statistical validation**
Finally, the translation is statistically validated upon data collection and the final adaption is made.
12) **Final version**
Based on the results of the statistical analyses, the translation is finalized.

Instructions for translators

When translating items to a new target language, translators are given the following instructions:

- The translation should follow the original wording as closely as possible. Adaptation is only allowed if necessary for making the text comprehensible, meaningful, or accurate in the target language. It can, however, be necessary to make slight adaptations to make the item idiomatic to the native speaking test person.
- The text should not be changed unnecessarily – especially, negations must not be removed or changed if this alters the coding or “direction” of the answers.
- Phrases, idioms, or other kinds of figurative speech should be omitted, as these may be unknown to or interpreted differently by different test persons.
- **Source versions**: Ideally, the original version (typical Swedish or Danish) or the English version or a combination of the two versions are used for translation.
- **Forms of address**: By default, the form of address should be informal, unless this is considered unprofessional, rude, or offensive (or in opposition to common usage) in the target language/culture.
- By default, forms of address should be neutral, unless it opposes common usage or is not a viable alternative in the target language.
- Non-gender specific terms are used, or feminine endings are added (e.g. in parentheses) whenever possible. If this is not possible in the language in question, our subject matter experts are involved in the decision.

9. References

Age Discrimination in Employment Act of 1967. Pub. L. No Pub. L. No 90-202, et seq (1967).

Allport, G. W. & Odbert, H. S. (1936). Trait names: A psycho-lexical study. Psychological Monographs, 47 (211), 171.

Barrick, M. R. & Mount, M. K. & Judge, T. A. (2001). Personality and job performance at the beginning of the new millennium: What do we know and where do we go next? International Journal of selection and Assessment, 9, 9–30.

Barrick, M. R. & Mount, M. K. (2005). Yes, Personality Matters: Moving on to More Important Matters. Human Performance 18 (4), pp. 359-372

Björnsson, C. H. (1968). Läsbarhet. Stockholm: Liber.

Borsboom, D. (2006). The attack of the psychometricians. Psykometrika, 71, 425–440.

Chissom, B. S. (1970). Interpretation of the Kurtosis Statistic. The American Statistician, 24(4), 19-22.

Christensen, K. B., Makransky, G. and Horton, M. C. (2017). Critical Values for Yen's Q3:Identification of Local Dependence in the Rasch model using Residual Correlations. Applied Psychological Measurement, 41(3), 178-194.

Conoly, C. A. (2003). Differential Item Functioning in the Peabody Picture Vocabulary Test – Third Edition: Partial Correlation versus Expert Judgment. Doctoral Dissertation: Texas A&M University.

Costa, P. T. & McCrae, R. R. (1982). An approach to the attribution of aging, period, and cohort effects. Psychological Bulletin, 92, 235–250.

Costa, P.T. and McCrae, R.R. (2008) The Revised NEO Personality Inventory (NEO-PI-R). In: Boyle, G.J., Matthews, G. and Saklofske, D.H., Eds., The SAGE Handbook of Personality Theory and Assessment: Volume 2—Personality Measurement and Testing, SAGE Publications, 179-198.

Cuppello, S., Treglown, L., Furnham, A. (2023). Personality and management level: Traits that get you not the top. Personality and Individual Differences, 206, 1-5.

Digman, J. M. (1997). Higher-order factors of the Big Five. Journal of Personality and Social Psychology, 73, 1246–1256.

European Federation of Psychologists' Associations [EFPA] (2013). Review model for the description and evaluation of psychological and educational tests. ver. 4.2.6. EFPA Board of Assessment.

Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U. S. D. of J. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38290–38309.

He, S. L. & Wang, J. H. (2012). Development of the Chinese version of the Dentine Hypersensitivity Experience Questionnaire (DHEQ). European Journal of Oral Science, 22, 218–23.

Howard, M. C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? International Journal of Human-Computer Interaction, 32(1), 51-62.

Hunter, J. E. & Schmidt, F. L. (2004). Methods of meta-analysis. CA: Sage Publications.

ISO10667 (2010). Bedömningstjänster i arbetslivet – Processer och metoder för bedömning av människor i organisationer. Del 1: Krav på uppdragsgivare. Del 2: Krav på leverantör, 2010 (No 10667.1.). Stockholm: Author.

Judge, T. A., Bono, J. E., Ilies, R. & Gerhardt, M. W. (2002). Personality and Leadership: A Qualitative Review. Journal of Applied Psychology, 87 (4), 765–780.

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.

Kang, W., Guzman, K. L., Malvaso, A. (2023). Big Five personality traits in the workplace: Investigating personality differences between employees, supervisors, managers, and entrepreneurs. Front. Psychol., 14, 1-8.

Kim, H. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Open lecture on statistics.

Mabon, H. (2005). Arbetspsykologisk testing. Om urvalsmetoder i arbetslivet. Stockholm: Assessio International AB.

Marais, I. (2012). Local Dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), Rasch Models in Health. Wiley-ISTE.

Mischel, W. (1968). Personality and assessment. New York: Academic.

Mussel, P., Winter, C., Gelléri, P., & Schuler, H. (2011). Explicating the Openness to Experience Construct and its Subdimensions and Facets in a Work Setting. International Journal of Selection and Assessment, 19(2), 145-156.

Roberts, B. W., Walton, K. E. & Viechtbauer, W. (2006). Patterns of Mean-Level Change in Personality Traits Across the Life Course. Psychological Bulletin, 132 (1), 1-25.

Rouquette, A., Hardouin, J. B., Vanhaesebrouck, A., Sébille, V., & Coste, J. (2019). Differential Item Functioning (DIF) in composite health measurement scale: Recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. PLoS ONE 14(4), 1-16.

Sackett, P. R., Zhang, C., Berry, C. M., & 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), pp. 2040–2068.

Schmidt, F. L., Shaffer. J. A. & Oh, I. S. (2008). Increased accuracy for range restriction corrections: implications for the role of personality and general mental ability in job and training performance. Personell Psychology, 61, 827–868.

Shen, L., Zeng, H. Zeng, H., Jin, X. Yang, J., Shang, S., & Zhang, Y. (2018). An Innovative Evaluation in Fundamental Nursing Curriculum for Novice Nursing Students: An Observational Research. Journal of Professional Nursing, 34(5), 412-416.

Smits, I. A. , Dolan, C. V. , Vorst, H. C. , Wicherts, J. M. & Timmerman, M. E. (2011). Cohort Differences in Big Five Personality Factors Over a Period of 25 Years. Journal of Personality and Social Psychology, 100 (6), 1124-1138.

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? Journal of Applied Psychology, 95(4), 781–790.

Staufenbiel, T. & Hartz, C. (2000). Organizational Citizenship Behavior: Entwicklung und erste Validierung eines Meßinstruments. Diagnostica, 46(2), 73-83.

Streiner, D. L., & Kottner, J. (2014). Recommendations for reporting the results of studies of instrument and scale development and testing. Journal of Advanced Nursing 70(9), 1970–1979.

Stricker, L. J. (1982). Identifying Test Items That Perform Differentially in Population Subgroups: A Partial Correlation Index. Applied Psychological Measurement, 6(3), 261-273.

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education, 48, 1273-1296.

Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). Personnel Psychology, 60(4), pp. 967–993.

Tupes, E. C. & Christal, R. E. (1961, 1992). Recurrent personality factors based on trait ratings. Journal of Personality, 60, 225–251.

Van Bork, R., Grasman, R. P. P. P., Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. Psychometrika, 83, 443-452.

Viswesvaran, C., Ones, D. S. & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81, 557–574.

Woods, C. M. (2009). Testing for Differential Item Functioning with Measures of Partial Association. Applied Psychological Measurement, 33(7), 538-554.

Ziljmans. E. A. O., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2019). Item-Score Reliability as a Selection Tool in Test Construction. Frontiers in Psychology, 9, 1-12.

Appendix A: MAP Essence

Due to the popularity of MAP in the market and a need for a shorter, more efficient assessment of personality, it was decided to develop a short personality test based on MAP. This product is labeled MAP Essence (abbreviated MAP-E). The aim of the development of MAP-E was to provide a reliable and valid assessment of personality at the FFM-level based on MAP but with a shorter completion time for the respondent.

MAP-E may be used as a single assessment or as part of a process with several assessments set up by the test administrator. For predictive purposes, a set of screening lenses for Integrity and Service are also available. MAP-E provides a brief but comprehensive measure of the five overall personality traits (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness) using a subset of 75 items included in MAP.

Development
The items in MAP-E are a selection of items from the 200 items in MAP. This implies that scores on Essence may be interpreted as indicators of scores on MAP. Note that the scales scores in MAP-E are somewhat less reliable and valid compared to scales on the full MAP due to the reduced number of items.

Sample
The sample used for development consisted of 25 733 respondents selected based on convenience sampling. Data was collected via Assessio's web-based test platform Assessio Select. The sample comprised 48 % females and 52 % males with a mean age of 37.3 years (SD = 10.9).

Scale construction
Items were selected using both theoretical and empirical criteria. The requirements were stated before the final selection of items was made. These requirements were:

- Each item content should (theoretically) reflect the trait in question.
- Each item should correlate most highly with the scale to which it belongs.
- Item content should not overlap theoretically.

The overall goal with the development of MAP-E was to develop a short assessment, preferably taking half the time completing compared to MAP. Therefore, it was decided to aim for less than 100 items to get a completion time of approximately 15 minutes or less.

First, with the aim of ensuring content and construct validity, it was decided to rely on the facet structure in MAP, i.e., making each facet contribute equally to the overall trait score (as opposed to using item-total correlations on the total trait score at the expense of coverage of the whole construct). To identify the most suitable items from the facets, item-total correlations between each single item and the facet were computed. Based on these analyses, three items from each of the facets were selected, thus totaling 15 items per trait and 75 items in total. At this point, a qualitative review was conducted to ensure that item content did not overlap theoretically. As a result of this review, three items were replaced by the second-best item in those facets due to overlap with other items.

Construct validity

To further support the construct validity, analyses exploring both convergent and discriminant validity of the scales were conducted. Table A1 shows the correlations between MAP-E scales and trait scores based on full MAP. Numbers on the diagonal (highlighted in bold) represent convergent validity (correlations with target, similar constructs), whereas off-diagonal numbers represent discriminant validity (correlations with unrelated or less similar constructs). On average, MAP-E scales had a correlation of .93 with target trait scores in MAP and a mean correlation of .35 with non-target trait scores (ranging from .29 for OP to .42 for CO).

Table A1. Convergent and discriminant validity of MAP-E scales.

MAP-E scale	EX	AG	CO	ES	OP
Extraversion	**.92**	.37	.39	.37	.35
Agreeableness	.30	**.93**	.41	.38	.28
Conscientiousness	.33	.57	**.90**	.55	.21
Emotional Stability	.32	.45	.54	**.95**	.04
Openness	.46	.32	.26	.12	**.93**

To test the overall model, a Confirmatory Factor Analysis (CFA) was carried out. The results of the analysis showed a statistically significant difference between the model and the data (Chi-square (569) = 60289.9, *p* < .001). Measurements of personality traits, however, seldom reveal non-significant chi-square values (which is often regarded as an overly conservative and unrealistic measure of fit between model and data), which partly depends on the fact that the chi square is influenced by sample size. Therefore, adjustment measures such as the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI) were considered. These indicated that the model needs some improvement to reach fully acceptable levels (RMSEA =. 09; CFI =. 69).

Reliability

The reliability of MAP-E scales is presented below in Table A2 with estimates of Standard Error of Measurement (SEM) at both the raw score and C score level fixed (standard deviation of 2).

All scales had sufficient reliability with alphas ranging from .70 to .83 with an average of .76. On average, scales had a Standard Error of Measurement (SEM) of less than 1 C score point.

Table A2. Reliability of MAP-E scales.

Scale	Items	Alpha	SD	SEM	SEM (C score)
Extraversion	15	.77	5.14	2.47	0.96
Agreeableness	15	.70	4.83	2.65	1.10
Conscientiousness	15	.77	4.49	2.15	0.96
Emotional Stability	15	.83	5.68	2.34	0.82
Openness	15	.73	5.23	2.72	1.04
Mean	**15**	**.76**	**5.07**	**2.46**	**0.98**
Median	**15**	**.77**	**5.14**	**2.47**	**0.96**

Norm group

In 2024, a new global norm was implemented replacing previous norms. The global norm consists of 20 053 people who were selected through stratified randomization from a total of 225,573 people aged 18-70 who completed the assessment in a high-stake setting (selection or development). Statistical analyses confirmed that the norm group does not represent a biased sample, as score differences between different samples were only small or negligible across scales (Cohen's d ranging from 0.09-0.46 with an average of 0.21). Hence, the norm group complies with EFPA standards for excellent sample size (1,000+) and update (less than 10 years old).

With regards to composition, the sample was stratified for gender at the nationality level, hence making each nationality contribute an equal number of main genders (M and F). Then, nationalities were stratified such that each nationality constituted a maximum of 2.5 % of the total norm group. Next, other genders were added to the norm group as well with the aim of having a representation of roughly 1 % but with the restriction that this addition did not cause any nationality to be largely overrepresented. Although the final age distribution was slightly skewed to the left (with a median of 32), statistical analyses revealed no significant relationships between age and any of the scores as evidenced by very low correlations ranging from -.11 to .07, with absolute values averaging just .10. Therefore, it was deemed unnecessary to stratify for age, as this would only reduce the sample size without impacting overall scores across age groups. As the final sample comprised a proper range of education levels and occupations (job families), and there were no major score differences, the sample was not further stratified for any of these demographic variables.

A brief overview of the demographic composition of the norm group is outlined in Table A3.

Table A3. Composition of the global norm group for MAP Essence.

MAP Essence: Global norm	
Update	2024
Data collection	2019-2023
Sample size	20,053
Composition	
Purpose	Selection: 73 % Development: 27 %
Gender	49.8% females 49.8% males .4% other
Age	18-70 (M = 33.7, SD = 9.37)
Nationalities	162, Max. = 2.54 %
Education level (%)	
Elementary school	2.7
Middle/Junior high/High school	21.7
Less than 3 years of post-secondary education	12.9
3 or more years of post-secondary education	55.9
PhD	2.5
Other	4.0
N/A	.3

Group differences

Finally, to further substantiate the quality of the norm group composition, group differences were examined with independent samples t-tests for test purpose (selection/development), gender (male/female) and age (dichotomized using 40 as the cut-off value) listed below in Tables A4-A6. The analyses showed that effect sizes (Cohen's d) were small (< .50) or negligible (< 0.20) for most scales, ranging from 0.02-0.27 across scales and demographic variables, averaging 0.16, 0.11, and 0.15 for purpose, gender, and age, respectively.

In conclusion, the global norm for MAP Essence constitutes a large, well-composed sample that is suitable for all intended target groups and applications (selection and development).

Table A4. Group differences on MAP-E scales for test purpose (selection/development).

Scale	Selection			Development			Comparison			
	N	M	SD	N	M	SD	Dif.	t	p	d
Extraversion	14,585	45.2	5.19	5,468	45.1	5.03	0.08	0.99	0.321	0.02
Agreeableness	14,585	49.1	4.75	5,468	47.8	4.90	1.32	17.32	< .001	0.27
Conscientiousness	14,585	53.2	4.39	5,468	52.3	4.67	0.92	12.97	< .001	0.21
Emotional Stability	14,585	49.3	5.52	5,468	47.9	5.97	1.40	15.62	< .001	0.25
Openness	14,585	46.5	5.23	5,468	46.1	5.23	0.40	4.78	< .001	0.08

Table A5. Group differences on MAP-E scales for gender (female/male).

Scale	Female			Male			Comparison			
	N	M	SD	N	M	SD	Dif.	t	p	d
Extraversion	9,982	45.4	5.06	9,982	45.0	5.22	0.34	4.74	< .001	0.07
Agreeableness	9,982	49.3	4.63	9,982	48.2	4.95	1.06	15.66	< .001	0.22
Conscientiousness	9,982	53.2	4.32	9,982	52.7	4.61	0.49	7.73	< .001	0.11
Emotional Stability	9,982	48.8	5.57	9,982	49.1	5.75	-0.28	-3.53	< .001	0.05
Openness	9,982	46.7	5.19	9,982	46.1	5.25	0.55	7.45	< .001	0.11

Table A6. Group differences on MAP-E scales for age (above/below 40).

Scale	40+			< 40			Comparison			
	N	M	SD	N	M	SD	Dif.	t	p	d
Extraversion	4,839	44.6	5.11	15,214	45.4	5.14	-0.80	4.74	< .001	0.16
Agreeableness	4,839	49.1	4.80	15,214	48.7	4.84	0.42	15.66	< .001	0.09
Conscientiousness	4,839	52.2	4.63	15,214	53.2	4.42	-0.95	7.73	< .001	0.21
Emotional Stability	4,839	49.3	5.46	15,214	48.9	5.74	0.45	-3.53	< .001	0.08
Openness	4,839	45.5	5.24	15,214	46.7	5.20	-1.21	7.45	< .001	0.23

Adverse Impact

In line with the analyses for MAP, Adverse Impact simulations were conducted for each of the MAP-E scales using strict, moderate, and lenient selection ratios (i.e., selecting the top 23, 40, and 60 %, respectively). Results for gender (female/male) and age (40+/<40) are displayed below in Tables A7 and A8, respectively.

For gender, no adverse impact is expected even when using strict selection ratios, except for Agreeableness that slightly favor females (unless moderate or lenient selection ratios are applied). For age, strict selection ratios should be avoided for Extraversion, Conscientiousness and Openness, the latter of which also showed a disproportionate hiring rate in favor of younger individuals at moderate selection ratios (AI ratio of .79).

Table A7. Adverse Impact simulations for gender (female/male) at different selection ratios (SR).

Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
Extraversion	0.34	0.07	0.93	0.95	0.96
Agreeableness	1.06	0.22	0.78	0.84	0.87
Conscientiousness	0.49	0.11	0.90	0.90	0.94
Emotional Stability	-0.28	0.05	0.89	0.93	0.96
Openness	0.55	0.11	0.86	0.91	0.94

Table A8. Adverse Impact simulations for age (above/below 40) at different selection ratios (SR).

Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
Extraversion	-0.80	0.16	0.78	0.84	0.92
Agreeableness	0.42	0.09	0.88	0.94	0.95
Conscientiousness	-0.95	0.21	0.75	0.80	0.88
Emotional Stability	0.45	0.08	0.94	0.96	0.97
Openness	-1.21	0.23	0.71	0.79	0.83

In conclusion, when applying proper selections ratios, MAP Essence provides a fair and unbiased assessment that does not cause any Adverse Impact for any protected group when used for making employment decisions.