Logics

Technical manual
Version 2.0

HELENE HOPPE REVALD
Director of Psychometrics, Occupational Psychologist

NANNA BAK SKYTTE
Lead Psychometrician, Occupational Psychologist

DENNIS HVASS
Psychometrician, Occupational Psychologist

Assessio

Table of contents

1. Introduction and Theoretical Background

Classical intelligence theory is rooted in the idea that cognitive ability can be measured and organized according to identifiable factors. One of the foundational contributions to this field came from Charles Spearman, who, in the early twentieth century, introduced the concept of a general intelligence factor - commonly referred to as the "g factor." Spearman's research revealed that individuals who perform well in one area of cognitive testing are likely to perform well in others, suggesting the presence of a broad, underlying mental capability that influences performance across diverse intellectual tasks.

However, Spearman also acknowledged the presence of specific abilities, known as "s factors". These are unique to particular tasks or domains and operate alongside the general factor. Later, Raymond Cattell refined the understanding of intelligence by distinguishing between two broad components: fluid and crystallized intelligence. Fluid intelligence encompasses the capacity to solve novel problems, reason abstractly, and adapt to new situations independent of prior knowledge. In contrast, crystallized intelligence represents the knowledge and skills acquired through experience and education, such as vocabulary and factual knowledge. Interestingly, fluid intelligence tends to decline with age whereas crystallized intelligence tends to remain stable or increase throughout the lifespan (Ryan et al., 2000).

Together, these theories form the backbone of classical intelligence theory, providing a framework for how general and specific cognitive abilities interact and contribute to overall intellectual performance. These foundational ideas continue to shape contemporary approaches to understanding and measuring human intelligence.

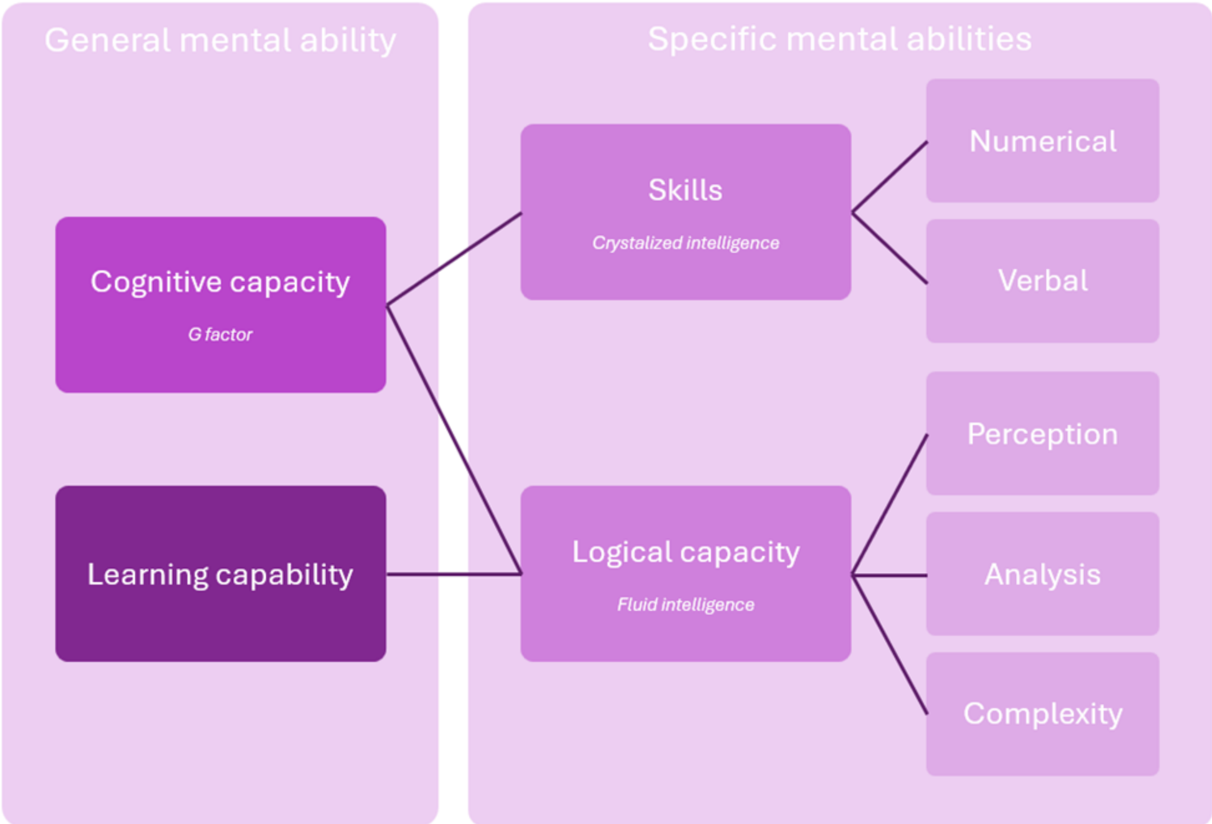
In the business context, intelligence tests are widely recognized for their ability to predict job performance and workplace success. General Mental Ability has consistently been shown to be one of the strongest predictors of training outcomes, problem-solving capacity, and overall job performance across a wide range of roles and industries (Schmidt et al., 2016). By assessing intellectual capabilities such as reasoning, comprehension, and problem-solving, organizations are better able to identify candidates who are likely to learn quickly, adapt to new challenges, and make effective decisions. As a result, intelligence testing has become an integral tool in personnel selection, talent development, and succession planning, helping companies build high-performing and adaptable workforces.

While classical intelligence theory emphasizes General Mental Ability, alternative perspectives have also emerged. Most notably, Howard Gardner's theory of multiple intelligences challenges the notion of a single, overarching intelligence. Gardner proposed that intelligence is better understood as a set of distinct capacities, such as linguistic, logical-mathematical, spatial, musical, and interpersonal abilities, among others. According to this view, individuals may excel in some areas while demonstrating average or below-average ability in others. This broader conceptualization of intelligence has influenced educational practices and expanded the dialogue around cognitive strengths and diversity without showing as much promise in their ability to predict job performance and workplace success compared to traditional intelligence measures.

Description of the model behind Logics

Like other well-documented, recognized intelligence tests within both business and clinical fields (e.g. Ravens Progressive Matrices, Wechsler Adult Intelligence Scale, WAIS), Logics was developed on the basis of classical intelligence theory, in particular Charles Spearman's distinction between a general (g) factor and specific (s) factors, as well as Raymond Cattell's distinction between fluid and crystallized intelligence. The various parameters in Logics thus constitute a broad and relevant selection of the aspects that define intelligence according to many years of international research. The test also provides detailed information about the test person's learning capability, efficiency, and decision style, which makes Logics particularly relevant in a business context.

The hierarchical structure of parameters in Logics and how they relate to classical intelligence theory is illustrated below. These hierarchical layers (or strata) mirror the Cattell-Horn-Carroll Theory of Cognitive Abilities, which has received extensive empirical support throughout several years of research (Flanagan & Dixon, 2014). In addition, the parameters in Decision style (Speed and Accuracy) and the GMA score are calculated based on the full set of items, and learning capability is based on a subset of items within Logical capacity with Fluid intelligence as the basis for future learning.



The concept of the "g factor" in Logics is thus expressed by an overall GMA score, referred to as Cognitive capacity.

In Logics, fluid intelligence is captured by the parameters under Logical capacity, divided into the ability to perceive and observe logically (Perception), the ability to think and deduce logically (Analysis) and the ability to comprehend larger amounts of complex information (Complexity). Crystalized Intelligence is located in Skills and divided into Numerical and Verbal abilities.

Initial development of Logics

The initial development of Logics was completed in the year 2000 as a combined intelligence and skill set analysis – with recruitment and personal development as its primary applications. From the very beginning, the requirements for the test were that it should have a good correlation with existing intelligence tests such as Raven and WAIS, and it had to be relevant to a business context. Therefore, part of the development process included a study of the needs of 50 Danish companies in connection with intelligence testing.

From the beginning, Logics sought to combine the two major traditions within classical intelligence theory and address the question of whether there is one or several types of intelligence. Therefore, the overall GMA score is generated, while at the same time, several sub-parameters inspired by tests like WAIS (Wechsler Adult Intelligence Scale) are measured. Logics does not attempt to measure all kinds of intelligence, but the classic understanding of the concept of intelligence, namely the general factor of intelligence (g factor) nuanced by specific mental abilities (s factors) within the domain of both fluid and crystallized intelligence.

Continuous updates to Logics

The scales, norms, and items are monitored and updated on a regular basis to ensure test quality. Examples of recent changes over the past years include:

- Updates to the wording of items to match modern-day, inclusive language (e.g., changing formal to informal forms of address or “men” to “persons”).
- Increases in item difficulty for items that have become too easy over time.
- Adding and revising items in the Itempool based on monitoring of item equivalence in terms of frequency, difficulty, and time use (see below).
- Ongoing additions and revisions of languages (functional testing) on a need basis.
- Updates to the algorithms measuring Cognitive Capacity and Learning Capability based on data to increase overall test quality and a more advantageous weight between speed and accuracy.
- Norm updates (as outlined in the section on Standardization).

Scale overview

The development resulted in a nuanced intelligence and skill set analysis consisting of 80 items (i.e., tasks) that form the basis for an overall score for GMA (General Mental Ability, referred to as Cognitive capacity), as well as 8 parameters (scales) that elaborate on more specific aspects of the test person’s specific mental abilities as well as decision style, skills, logical capacity, and learning capability.

The parameters in Logical capacity and Skills have their own set of items. The parameters in Learning capability are based on a subset of items from Logical capacity. The parameters within Decision style and Cognitive Capacity reflect the entire set of tasks and are thus based on all items.

2. Scale Definitions and Interpretations

The following chapter outlines the main definitions of the 9 parameters (scales) in Logics, organized into five major domains of Cognitive capacity (GMA), Learning capability, Decision style, Logical capacity, Skills.

General Mental Ability

Cognitive Capacity

A measure of the general mental ability/overall intelligence, which examines the overall cognitive capacity compared to a norm. The GMA score is calculated on the basis of all tasks in the test and is a complex function of the test person's speed and accuracy in their completion of the test. The GMA score includes aspects of both fluid and crystallized intelligence, weighted equally. The score reflects the underlying factor that can explain the individual's ability to quickly and precisely solve and learn how to solve different cognitive tasks of different familiarity (known as the g factor).

People with high scores tend to quickly and correctly cope with abstract, numerical and verbal reasoning tasks that require logical processing and problem-solving of comprehensive and faceted information. They are likely to handle even excessively challenging cognitive tasks successfully. They are able to solve most types of cognitive tasks but can get bored if they are too easy.

Individuals with low scores tend to find several types of at least some cognitive tasks slightly difficult. They typically need more time, introduction and experience in order become familiar with cognitively challenging tasks. They may struggle to process too many and too large cognitive demands and therefore often need thorough instruction or simpler tasks.

Learning Capability

Learning capability is a measure of the ability to learn new things given the necessary time, and it examines how accurately novel logical problems are solved (in the first half of the test). The measure of Learning capability is based on tasks from Logical Capacity, which measures fluid intelligence and examines the person's capability to solve novel logical problems without any prior knowledge, regardless of whether the person has had enough time to see and solve all the cognitive tasks in the test.

People with high scores easily learn new things and solve novel logical problems. They are highly capable of acquiring new knowledge if they are given the necessary time. They understand, analyze and handle complex information well and therefore often seem like quick learners but they may need more time to complete their tasks (depending on the time used).

Individuals with low scores may benefit from introduction and supervision when learning new things. They may need repetition and additional time to acquire new knowledge. They may find it difficult to understand and analyze novel logical information, especially if there are too many inputs, and therefore often need sub-tasks with thorough instructions in both content and methods.

Decision style

Information about a person's problem-solving strategies based on the prioritization of speed and accuracy in the person's completion of the entire logical test.

Speed

Speed is a measure of pace, which examines how fast the person draws conclusions and makes decisions when presented with cognitive tasks. Speed measures what percentage of tasks the person has answered in total. The result on Speed must be considered in relation to the score on Accuracy.

People with high scores quickly come to conclusions when presented with cognitive tasks. They are fast in thought and action and generally rapid decision makers. They may seem bright and as quick thinkers, but there is a risk that they will spend too little time thinking things through and thus making rash conclusions.

Individuals with low scores take time to solve cognitively challenging problems. They tend to be slow-working and have difficulty in making quick decisions. They may seem thorough, as they often think things through before they make a decision, but there is a risk that they will lack efficiency.

Accuracy

Accuracy is a measure of correctness, which examines how accurate the person is in their conclusions on cognitive tasks. Accuracy measures what percentage of the person's answered tasks that have been answered correctly. The results must be considered in relation to the score on Speed.

Individuals with high scores draw the right conclusions on the chosen cognitive tasks. They are thorough in their problem solving and quality-focused in their work. They typically deliver high quality and set high standards, but there is a risk that they will become too perfectionistic in their work.

People with low scores tend to reach inaccurate conclusions when solving cognitive problems. They often make many mistakes and may have a superficial way of working. They may lack knowledge or understanding of the tasks, but they may also tend to guess or answer without deeper consideration.

Logical capacity

Information about a person's fluid intelligence based on abstract reasoning tasks that examine the ability to understand, analyze and handle complex information when exposed to novel logical problems.

Perception

Perception is a measure of logical comprehension, which examines the person's ability to recognize patterns and see logical connections. The tasks that measure Perception are designed to be difficult to understand but easy to solve and they mainly consist of non-verbal figures and illustrations. Therefore, Perception is the one of the five areas which is least dependent on language abilities.

People with high scores recognize and understand abstract patterns with ease. They see connections in a larger context and easily comprehend logical tasks. They will typically perceive things correctly from the start, but they may find it difficult to accept if others do not.

Individuals with low scores have some challenges perceiving abstract patterns. They may overlook key elements and sometimes struggle to comprehend logical tasks. They may not always get the input right and therefore tend to rely more on subjective interpretations or ask many questions to understand the task.

Analysis

Analysis is a measure of logical reasoning, which examines the person's ability to think rationally and process logical information in a systematic way. The tasks that measure Analysis are designed to be easy to understand but difficult to solve and they are primarily based on a verbal source of information.

Individuals with high scores conclude rationally and logically on the available data. They often approach their tasks methodically and process logical information in a systematic way. They can generally work independently once the task is understood.

People with low scores risk drawing inaccurate conclusions from the available information. They tend to struggle processing logical information in a systematic manner. They may find it difficult to apply consistent methods and therefore often need guidance in tasks requiring logical thinking.

Complexity

Complexity is a measure of the ability to handle complex information, and it examines how a person uses logical deductions to comprehend and navigate in complicated logical tasks. The tasks that measure Complexity all holds a lot of information that the person need to keep track of at the same time. Therefore, Complexity is the one of the five areas which is most time consuming to complete.

People with high scores are able to navigate and solve complex logical problems. They have a good overview and can handle a lot of complicated information at once. They often prefer tasks requiring advanced problem-solving or handling multiple competing demands, but they may have a tendency to overcomplicate things.

Individuals with low scores have some difficulty processing multiple information simultaneously. They easily lose track and can have difficulty solving very complicated tasks. They quickly become overwhelmed by too many inputs and often tend to prefer simpler tasks.

Skills

Information about a person's crystallized intelligence that examines verbal and numerical reasoning skills based on tasks, which require acquired information and prior knowledge to solve them properly.

Numerical

Numerical is a measure of numerical reasoning, which examines number sense and mathematical skills. The Numerical tasks include basic calculations and arithmetic problem-solving.

People with high scores have a strong sense of numbers, calculations and rules of arithmetic. They are often able to apply their mathematical skills in various situations. They can solve most mathematical problems by themselves but may find it difficult to accept mathematical errors.

Individuals with low scores can be challenged by mathematical tasks. They tend to struggle with even basic numerical tasks and have difficulty relating realistically to numbers. They often find it difficult to solve mathematical problems on their own and often need to rely on external aids like calculators to solve them.

Verbal

Verbal is a measure of verbal reasoning, which examines the understanding of language, grammar and linguistic abilities in the test completion language. The Verbal tasks are typically built on word comprehension and basic grammar.

Individuals with high scores have an excellent grasp of grammar, vocabulary, and word comprehension. They are often able to express themselves with precision and nuance. They generally have strong language skills but may find it difficult to abstract from linguistic mistakes and limited language proficiency.

People with low scores show limited language proficiency in the test completion language. They tend to make spelling, grammar or other linguistic mistakes. They tend to struggle understanding and applying the test completion language and misunderstandings may occur as a result from limited vocabulary or other language difficulties.

3. Instructions for Use

This chapter provides guidelines for administering, interpreting, and providing feedback on Logics. The purpose of the following instructions is to create optimal conditions for the test administrator to provide the test person with the opportunity to complete the assessment in a standardized way, thereby ensuring fair and comparable results.

Areas of use

In line with the ever-increasing demands placed on both managers and employees in relation to job performance, the complexity of tasks, and the speed of task execution, the need to assess and predict people's ability to meet these demands is also increasing. By applying logical tests, companies can focus on people's individual strengths and see how a person matches a particular job, tailor a position to the person and ensure that the employee's resources are used in the best possible way.

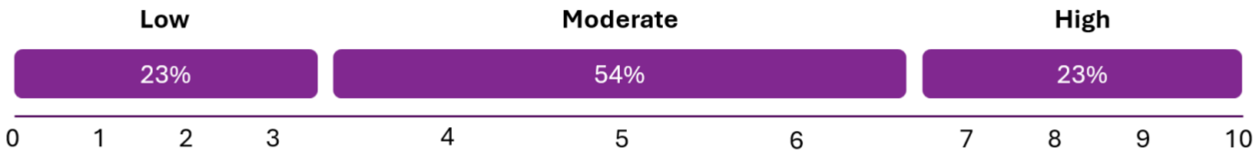
Primarily, Logics is used for recruitment and to provide information for management. At the same time, we consider it a management tool – the result should not simply be filed away but should instead be actively used in day-to-day management in relation to situations, individual management, learning styles, motivation, communication, and needs. Hence, the information in Logics can also be used for development, onboarding and succession planning.

Administration and scoring

Logics is built to be administered online via a link (unsupervised) but can also be used in an on-site setting (proctored) through the web-based platform. It consists of 80 items to be completed within a time limit of 30 minutes. Throughout the test, the test person can navigate freely between all items and choose which items to skip, answer and in which order. In addition to the answers, response times are also tracked to indicate the total time use as well as the average time used on items within specific parameters (scales). Throughout the test, a timer is available showing how much time is remaining.

In addition to an overall GMA score (Cognitive Capacity), Logics provides detailed scores from 0–10 on 8 additional parameters depicting the percentage of correct, incorrect answers, and skipped tasks, as well as average time used, depicted as a state (quick, average, slow) aimed at analyzing subdimensions or components of the GMA score to highlight individual strengths and weaknesses.

Each of the scales in Logics follow the C score distribution grouped into low, moderate and high as follows:



Requirements and conditions of testing

Environment

The test must be completed in one session without any distractions. To reduce the likelihood of cheating, the timer will keep going if the test person is interrupted. This is clearly stated in the instructions for the test.

Target group

Logics is suitable for employees aged 18-70 working at the salaried level and above. Although suitable for selection for a wide range of occupations and job industries, the Logics is especially useful for specialist and management positions (and positions of medium to high job complexity). Logics must not be used for any clinical, including diagnostic, purposes.

Computer skills

As Logics is administered digitally, basic computer skills and experience of working online are 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.

Information before testing

When a test person completes Logics, it is important that they know the purpose of the testing and the process in advance. Therefore, the test administrator should inform the test person about the following:

- How the process is conducted.
- When the test is to be completed.
- How the test is administered (online via link or on screen in a controlled setting).
- How the test results are used.
- How much weight is placed on the test results in the overall assessment.
- If, when and how feedback on test results is provided.

Before starting the assessment, the test person is given the following information:

- The assessment consists of 80 items.
- The assessment has a time limit of 30 minutes, upon which the test session ends.
- Only pen and paper are allowed as aids.
- Instructions should be read carefully before starting the assessment.
- That they should choose their native language if possible and that the language cannot be changed during the assessment.
- That they need to be able to complete the test without being disturbed.
- How their data will be processed, stored and for how long.

Before starting the test, the test person will be taken through an introduction to the assessment navigation as well as three practice items.

If the test person is invited to complete the assessment in a controlled setting, it is the sole responsibility of the test administrator to provide the test person with the necessary information, equipment, and conditions to complete the assessment and to take into account any adverse effects arising from completing the assessment in non-optimal conditions.

Adverse conditions

A basic level of reading comprehension is required to understand the instructions and items for Logics. Therefore, a range of disabilities or cognitive impairments might negatively impact overall results. As these are considered sensitive information according to GDPR principles, it is not possible to collect and/or display the information to the user. In addition, the exact impact on scores is difficult to estimate, as the type, number, and severity of different disabilities vary between individuals and might impact scores differently. However, some general trends and guidelines are presented below for different adverse conditions.

Dyslexia or other reading disabilities

Dyslexia is likely to impact and reduce overall reading speed (thus impacting Speed) as well as tasks requiring verbal understanding (Verbal) and/or tasks with substantial amounts of text to read (Numerical, Complexity). Depending on the severity of the condition and the number of items completed within these parameters, dyslexia might also have a slight impact on the overall GMA score.

Dyscalculia

For dyscalculia (difficulty in understanding and working with numbers), Numerical is the score most likely to be impacted. As dyscalculia can also affect visual processing and short-term memory, scores on Perception might also be affected.

Completion in a non-native language

Similarly, completing the test in a non-native language might reduce overall reading speed and word comprehension, thus impacting the scores on Speed, Verbal, Numerical, Complexity and to some extent GMA (depending on proficiency level in the completion language). In general, test persons are advised to choose their preferred language when starting the test, preferably their native language if available. Users are advised to take into account any negative effects of completing the test in a non-native or non-preferred language.

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 administering the assessment.

Prevention of cheating

The most frequent kind of cheating when completing the Logics is the use of aids such as calculators, dictionaries or AI. In addition, we are aware of some cases where candidates have had another person complete the test for them.

The optimal way to prevent these types of cheating is to complete the test in a controlled setting (i.e., supervised or proctored testing). This means that the test person is invited half an hour before their job interview and completes the test on a computer or tablet provided by the administrator. Before they begin, it should be emphasized that no aids other than pen and paper may be used.

If it is not possible to let the candidate complete the test under controlled conditions, we recommend the administrator includes questions around how the test was completed, as part of a job interview. This will allow the administrator to gain insight into the candidate's overall strategy and compare this with the result. For instance, if the test person is asked how many tasks the person managed to complete and the answer does not match the test result, this may give rise to a suspicion of cheating.

Occasionally, an opportunity to complete a test-like check of the result is offered if the test is completed at home. It is not possible to create a valid tool that can be completed in a short time and thus be easy and simple to use. In addition, accusing someone of cheating is a sensitive matter. Therefore, for ethical reasons, we have chosen not to offer this kind of service.

Recently, the use of LLMs (Large Language Models) or other AI-based technology has emerged as yet another tool that can be used for cheating (especially for Skills). At Assessio, we are constantly monitoring the performance and potential impact of various LLMs on our assessments. Analyses conducted each quarter from mid-2022 to mid-2025 show that scale means, medians, 93rd percentiles, and average time use on both numerical and verbal tasks are stable over time, suggesting that cheating with LLMs is not a widespread issue.

Guidelines for retesting

A well-known challenge when using GMA tests is that a person doing a retest can benefit from having seen the tasks before. We have addressed this by automatically replacing the task set during retesting (by means of an Itempool). The replacement of tasks works such that each test is composed of 80 tasks randomly selected from a larger pool of tasks. Hence, if the same test person completes a Logics again, they will not be shown the exact same 80 tasks as when they last completed it. The Itempool means that in the event of a retest, the test person is less likely to be able to take advantage of their memory.

As mentioned, the pool of tasks is developed at task level, which ensures that each test person is presented with exactly the same number of tasks of the same type and difficulty. This avoids candidates being given a set of very similar tasks, while other task types are completely missing (e.g., a person will not be presented with only Numerical tasks that contain percentage calculations and fractions, while not getting any tasks with other calculations, equations, problem solving, or number sequences). When developing new tasks, our rigorous testing and statistical analyses ensure that all tasks are identical in content, form and difficulty. We have set strict quality requirements to ensure that test persons can be compared irrespective of the specific item set, which is crucial for valid, fair, and ethical usage of the test.

In addition to the item pool, response options on each item are randomized to make the test less recognizable to individuals who complete it again. The randomization of response categories means that test subjects cannot rely on their visual memory when revisiting and answering tasks on Logics, because the placement of the response options is different between sessions.

Although these measures reduce the advantage of having completed Logics previously, retesting should be avoided or reduced as much as possible. Even though the test subject will not be shown the exact same 80 items at retesting, the person might still benefit from acquired knowledge of the structure, content, and form of the test. In addition, the test subject might choose a different strategy, either subconsciously or because of the previous strategy's impact on test results.

In some situations, however, retesting is warranted. In these situations, item replacement will enhance the validity of retesting. Technical difficulties (e.g., with the internet connection), errors in test completion, or incorrect choice of language are examples of such situations. In some cases, the test subject might also have experienced interruptions or pressure to such an extent that a retest is necessary. It should be noted, however, that a surprising or unexpected negative result on its own neither justifies nor entitles the test subject to a retest.

General principles for interpretation and feedback

Before the feedback session

Cognitive ability is a sensitive topic in many ways. Therefore, the individual may attach great significance to certain scores, and such scores may be surrounded by much frustration (especially low scores). Also, a certain amount of test anxiety is to be expected.

Therefore, the test person must be informed about the process and purpose of testing. However, the test administrator should be careful not to inform too much about the test before it is completed. The test person should be informed that the test must be completed in a quiet, undisturbed environment with no aids other than pen and paper. Furthermore, the test person should know that the test consists of 80 tasks and that it automatically shuts down after 30 minutes. The most important thing is to not give the test person guidelines for how to complete the test or any other information that can affect results.

When the test is completed, the test person gets feedback automatically with a printable overview of results for each of the scales. Before a potential feedback session or job interview, it is important to read through the test results and prepare relevant hypotheses and combinations that can be confirmed or dismissed through the dialogue with the test person during the feedback session.

During the feedback session

Feedback on results from Logics lasts approximately 30 minutes when provided in isolation. The special thing about giving feedback on a GMA test compared to personality or other assessments is that results reflect an actual performance as opposed to the test person's own judgement. Moreover, there is a tendency for the test person to be far more curious, excited, anxious, or worried about the results, which means that especially the beginning of the feedback session must be handled somewhat differently.

As an introduction to the feedback session, it can be advantageous to ask about the completion of the test before explicitly going through the results. That way, the administrator make sure that the feedback session will focus on questions and answers – not just the overall GMA score and other results. When going through the results, it is a good idea to begin with the GMA score to the test person. After that, you can explain the structure of the test and focus on the parameters.

The GMA score is a summation and a measurement of the test person's overall level. Therefore, it is not necessary to ask questions about it. Instead, present it in a neutral manner by telling where they are placed and put it into context by informing about the norm group.

After giving the GMA score, the structure of the test is explained to the test person. Sum up the following:

- What the remaining four areas cover (Learning capability, Decision style, Logical capacity, Skills).
- How the parameters are measured, their average and the score.
- What a low, moderate and high score means – present the norm group.
- What the legend shows – explain answer patterns and speed indicators.

Feedback on the parameters

There are different ways to present parameter scores. Either you can compare each specific score with the average of the norm group, or you can relate the different scores to each other. The latter means looking at the relative strengths and weaknesses of the candidate and lining them up against one another to identify contrasts, allowing you to focus on the best results of the person in question. This strategy is particularly useful when dealing with a person in a vulnerable situation, i.e., when one or more of the scores are low or the overall result is high.

It is important to point out that the use of either one or the other method does not need to be completely stringent; it is okay to use different approaches in the same session. Additionally, it is important to be aware that a candidate whose results suggest some degree of vulnerability does not necessarily find their results problematic. It is merely a matter of increased attention before and during the interview to ensure a good candidate experience with a non-judgmental approach.

Questioning technique

When giving feedback on the parameters, it is important to be aware of getting around the score and the interpretation of the parameter. Additionally, it is important to ask questions that:

- Are open and curious.
- Are not judgmental.
- Relates to the everyday work behavior of the test person.

Hence, the administrator needs to focus on their questioning technique to get as much information as possible from the feedback session. Excellent questioning technique is about asking questions that elaborate, explain and invites further dialogue. A rule of thumb is to use mainly open-ended questions, i.e., starting with “what”, “how”, “which”, “when”, etc.

There is no requirement that everything in the test must be covered equally thoroughly but everything should be mentioned. The most important thing is that the administrator, in their role as feedback provider, gain knowledge in the areas that are important in terms of the job and that the test person feels that they have been listened to and assessed fairly.

4. Scale Construction

The following section outlines the basic procedures for organizing and constructing the various parameters (scales) in Logics that are based on either the full set of 80 items or a subset of items.

Cognitive Capacity

In general, it is a well-known challenge for speed tests that speed can be given a disproportionate weight in the calculation of the overall score. In other words, this type of test risks favoring individuals who choose a strategy prioritizing speed (perhaps involving more guessing) as opposed to a more thorough strategy. Therefore, we have constructed a calculation model for Cognitive Capacity that gives uniform weighting to Speed and Accuracy (both of which are an expression of overall capacity), thus preventing favoring certain individuals based on their completion strategy. In addition, a few points are added if the test is completed in less than 30 minutes.

Table 4.1 shows the basic scale statistics for the sample used to construct the overall score alongside correlations with Speed, Accuracy and time spent completing the entire test. As expected, there is a strong positive correlation with both Speed and Accuracy as well as a small negative correlation with the total time used to complete the test.

Table 4.1. Scale statistics for the GMA score in Logics.

Scale statistic	Cognitive Capacity
N	3,537
Minimum (z)	-2.56
Maximum (z)	3.36
Skewness	0.32
Kurtosis	-0.24
Correlations	
Speed	.67
Accuracy	.72
Total time used	*-.07*

Learning Capability

The algorithm behind Learning was updated and simplified in 2025. The previous algorithm relied on learning sequences, i.e. measuring the progress on similar tasks in terms of increased speed (Efficiency improvement) and improved accuracy (Learning). However, this algorithm had a number of drawbacks:

- The measurement of efficiency was too dependent on initial speed (i.e., a high score was difficult to achieve when starting off quickly).
- In some cases, parameters could not be calculated because too many items were unseen or skipped (especially for items appearing in the second half of the test).
- The Learning measure correlated too strongly with overall Speed and Cognitive Capacity.

For these reasons, we decided to base the Learning capability entirely on items within the first half of the test. To minimize the impact of specific items not being seen or answered, learning sequences were also removed from the calculation and replaced with a larger, non-sequential subset of items. To capture the ability to solve new, unknown tasks independent of previously acquired knowledge, the item subset relies only on items from Logical capacity (i.e., fluid intelligence).

Finally, Efficiency improvement was removed as a separate scale score and replaced with a speed indicator indicating the relative time spent on items within this subset compared to the norm group average.

Table 4.2 shows the correlations of the new and previous parameters in Learning capability with selected Logics parameters. As expected from the principles outlined for the algorithm, the new Learning measure correlates strongly with Accuracy, the previous Learning parameter and GMA and weakly with Speed and the previous measure of Efficiency improvement. In conclusion, the correlations support the notion that the new scale is a measure of learning that is related but not identical to the score on Cognitive Capacity.

Table 4.2. Relevant correlations for new and previous parameters in Learning capability.

Scale	Efficiency imp.	Learning	New Learning scale
Speed	.36	.60	.14
Accuracy	.03	.54	.75
Efficiency imp.	-	.22	.05
Learning	.22	-	.59
Cognitive Capacity	.25	.81	.64

Decision style

The parameters in Decision style are calculated as the proportion of tasks solved relative to the full set (Speed) and the proportion of correct answers relative to the number of tasks solved (Accuracy).

Logical capacity and Skills

The parameters of Perception, Analysis, Complexity, Numerical and Verbal are calculated as the proportion of correct answers compared to the total number of items within each parameter (16 on average). The number of items for each parameter have been varied to consider the differing amount of time spent solving the items (for instance, Complexity has fewer items than Analysis) and to ensure equal weighting of fluid and crystallized intelligence in the total set of items.

Response patterns

In addition to the total score, these parameters are supplemented by information on the proportion of tasks answered correctly, incorrectly and tasks that were seen but not answered (skipped). In addition, the average time spent on each item within these parameters is shown and grouped into quick (bottom 30 %), average (middle 40 %), and slow (top 30 %). In other words, average time use is defined as being within ±0.52 z score points from the mean.

Itempool

In the spring of 2021, we implemented an Itempool on Logics. With this Itempool, the test consists of 80 items (tasks or questions) that are randomly selected from a larger pool (or bank) of items. Hence, if a test subject completes the test again, the 80 items will differ from the first completion. An Itempool thus reduces the advantage of recalling some items from memory when redoing the test.

Unlike others, the Itempool was constructed at the item level, which means that the pool of new items was developed by "cloning" or replicating each of the original 80 items. This method ensures that, irrespective of item randomization and composition, all test subjects complete the exact same number of item formats across the entire test as well as within the individual scales. One alternative is to construct the Itempool at the scale level, which suffers the drawback that some test subjects might be given too many items with similar content (e.g., numerical tasks containing only percentage calculation and fractions but no tasks with elementary arithmetic, equations, or number series). The item-level development also has the advantage that each test subject is given a unique version of the test, regardless of how many times the test has been completed before.

Construction of the item pool

All items were constructed by replicating each of the original items' structure, form, content, and wording. Subsequently, new items were systematically tested and subjected to statistical analysis based on a large amount of data (ideally at least 500 cases for each item).

Items were approved based on an overall evaluation of their similarity with the original items. If they did not meet quality requirements, items were discarded or revised and tested again. To achieve maximum quality, new items were developed, rated, and approved by at least two members of our own team of psychologists and psychometricians.

Quality requirements encompassed qualitative (content, structure, and wording) as well as quantitative (statistical) criteria (response distributions, item difficulty, agreement, and correlation), the latter of which are elaborated below. Only items satisfying all criteria were approved to enter the final Itempool.

New items are being added to the Itempool continuously, as they are developed and tested according to the procedure presented above. In addition, all existing and new items as well as the functioning of the entire Itempool are continuously maintained and monitored.

Statistical analyses

Following systematic testing and data collection, new items were compared to the original items using a range of statistical analyses. Comparisons emphasized aspects of item difficulty as well as correlations, assessed with different statistical tests.

Item difficulties for new and original items were compared using 3 different measures:

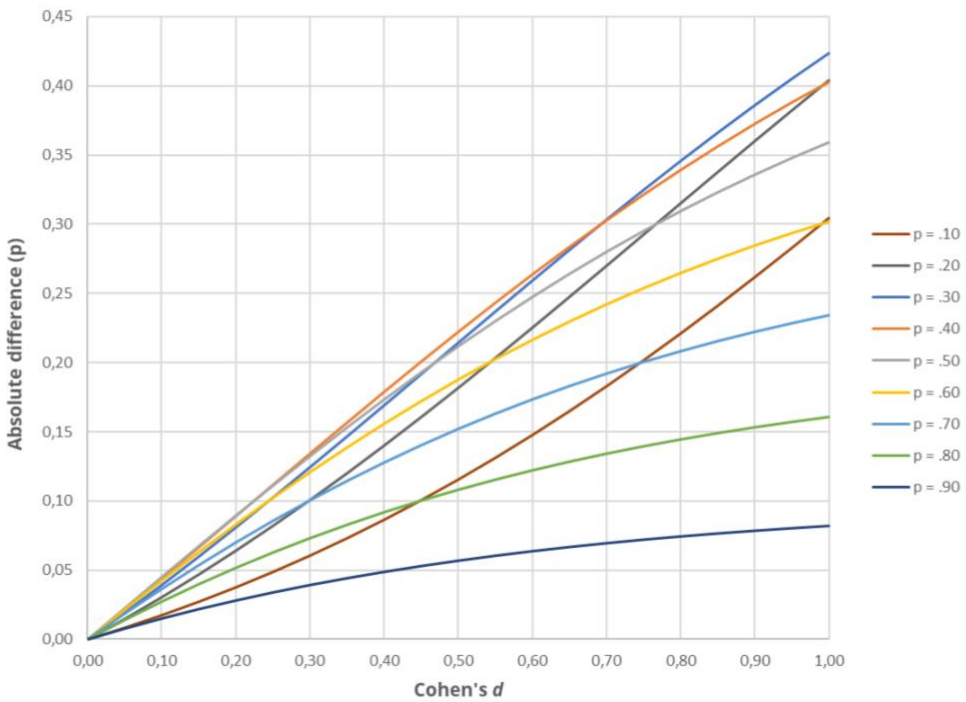
- The absolute **difference** in item difficulty were measured by comparing proportions of correct answers (%) between new and original items.
- A **McNemar test** was used to test if the proportions of correct answers (%) for new items differed significantly from original items.
- **Cohen's *d***: The difference in item difficulty was converted to a standardized effect size by converting the Odds Ratio (*OR*) to Cohen's *d* following the procedure outlined by Chinn (2000).

The *correlations* between new and original items were also examined in 3 different ways:

- **Agreement** was measured as the proportion of test subjects (%) who answered either correctly or incorrectly on both items.
- **Phi ratio**: The correlation between the new and original item was measured with the Phi coefficient. As Phi, unlike other correlation coefficients, cannot always obtain the maximum value of 1, it was converted to a ratio of the maximum value (*Phi ratio = Phi/Phi$_{max}$*).
- The ***p*-value** of the Phi coefficient was used to test if the correlation between the new and original item was statistically significant.

For each of these analyses, a range of criteria were set for maximal differences in item difficulty and minimal degrees of correlation. New items were always evaluated based on their overall performance on several criteria, as the different criteria are interdependent as well as dependent on the item difficulty in question. Hence, the maximally allowed difference in item difficulty is lower for easier as compared to more difficult items (and vice versa), illustrated below in Figure 4.1. In other words, more strict criteria were set for easier compared to more difficult items.

Figure 4.1. Difference a as a function of Cohen's d for different item difficulties.



Results of the statistical analyses are shown below for a standard performing item in each parameter in Table 4.3. All differences in item difficulty between new and original items were sufficiently low, and correlations were strong and statistically significant. In addition, results demonstrate the importance of using multiple criteria when assessing items. Although the difference in item difficulty appears larger for Analysis compared to Verbal, standardized effect sizes show the opposite pattern, because the two items does not have the same difficulties. Across all new items, the average difference in item difficulty was just 5 %-points with a small effect size of .18, on average. Overall, correlations were moderate to strong with an average of .37.

Table 4.3. Examples of statistical analysis of new and original items for each Logics parameter.

Parameter	*N*	Item difficulty			Correlations		
		Dif.	*p*	*d*	% Agr.	Phi ratio	Phi *p*
Perception	498	-.21	.815	0.07	96	.48	< .001
Analysis	278	12.16	.001	0.35	77	.39	< .001
Complexity	466	7.32	< .001	0.28	83	.57	< .001
Numerical	369	-2.28	.353	0.08	75	.34	< .001
Verbal	522	7.18	< .001	0.54	87	.41	< .001

Analyses of time use

As Logics is a speed test, it is crucial that the time spent on new items does not differ from original items, as this could affect overall test results. Therefore, average time spent (in seconds) on new and original items were compared using paired samples t-tests presented in Table 4.4.

Time spent was similar for new and original items, with standardized effect sizes being low to moderate. Across all new items, the average time used on new items differed by .87 seconds from time spent on original items. The absolute difference in time usage between new and original items averaged 7.0 seconds with an effect size of .39 on average.

Table 4.4. Statistical analyses of time spent on new and original items for each Logics parameter.

Parameter	*N*	Difference (s)		*t*	*p*	*d*
		M	*SD*			
Perception	498	3.31	9.50	7.77	< .001	0.35
Analysis	278	7.86	8.57	15.35	< .001	0.92
Complexity	466	3.68	13.38	5.93	< .001	0.28
Numerical	369	-6.86	36.91	-3.57	< .001	0.19
Verbal	522	3.38	8.80	8.80	< .001	0.38

Scale-level analyses

In addition to the item-level analyses presented above, scale-level analyses were conducted to examine how the Itempool affects overall scores and reliability of the individual scales (parameters). Table 4.5 shows the results of the analysis of the Numerical scale, in which an alternative version consisting of 44 % new items were compared to the original scale.

The inclusion of new items did not significantly affect the proportion of correct nor incorrect answers, tasks that were not seen/answered, the overall score, nor the reliability of the scale (including the Standard Error of Measurement, SEM, which is an estimate of the degree of uncertainty regarding the estimated score). Overall, this shows that the performance of new items is like that of original items, thus demonstrating scale equivalence.

Table 4.5. Comparison of the Numerical scale consisting of partially new or original items.

Score	Alternative	Original	Dif.	*z*	*t*	*p*	*d*
Correct	9.7	9.6	.2	-	0.98	.329	0.05
Incorrect	8.3	8.5	-.2	-	-0.98	.329	0.05
Not seen	3.4	3.2	.2	-	1.25	.213	0.07
Not answered	1.9	2.1	-.2	-	-1.64	.102	0.09
Overall score	54.1	53.1	1.1	-	0.98	.329	0.05
Reliability							
Alpha	.77	.79	-.02	-.93	-	.353	-
Split-half	.80	.82	-.02	-1.06	-	.290	-
SEM	1.69	1.68	.02	-	-	-	-

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.

In the development of Logics, it has been crucial that tasks have an appropriate difficulty level, giving the test person a good experience and making it pleasant and challenging for most people to complete the test. Therefore, Logics is designed as a speed test, characterized by tasks that are relatively easy to solve but challenge the person as the test is time-limited – rather than a power test that uses increasing difficulty as a means of examining the test subject's capacity, which can often be more time consuming and uncomfortable to complete, because the test typically ends when a series of incorrect answers is given.

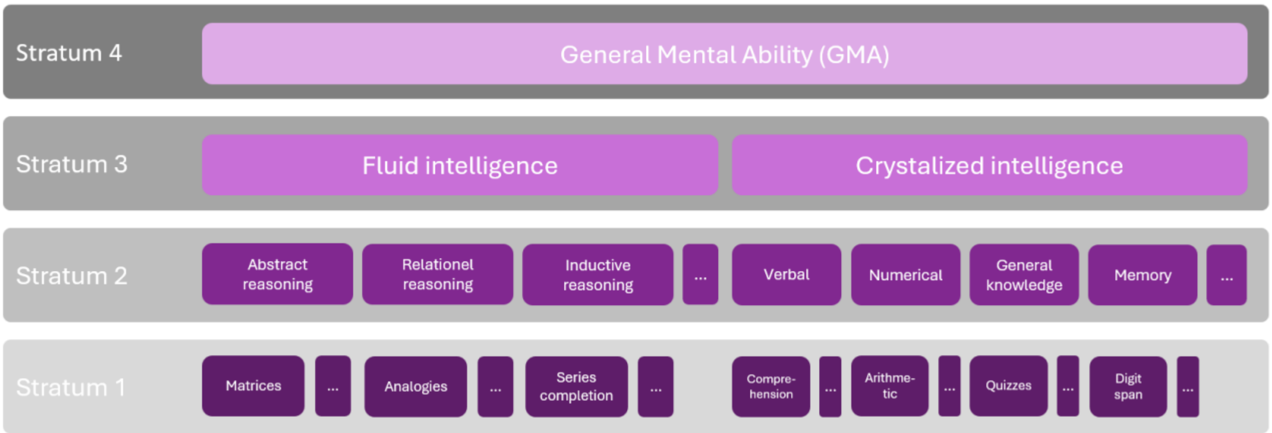
Content validity

As mentioned earlier, Logics has been developed on the basis of classical intelligence theory, in particular Charles Spearman's distinction between a general (*g*) factor and specific (*s*) factors, as well as Raymond Cattell's division of the concept of intelligence into fluid and crystallized intelligence. The various parameters in Logics thus constitute a broad and relevant selection of the aspects that define intelligence according to many years of international research. The test also provides detailed information about the test subject's learning capability, time use, and decision style, which makes Logics particularly relevant in a business context.

The hierarchical structure of parameters in Logics and how they relate to classical intelligence theory is shown below in Figure 5.1. These hierarchical layers (or strata) mirror the Cattell-Horn-Carroll Theory of Cognitive Abilities, which has received extensive empirical support throughout several years of research (Flanagan & Dixon, 2014). In addition, the parameters in Decision style (Speed and Accuracy) are calculated based on the full set of items (like the GMA score), and learning capability is based on a subset of items within Logical capacity.

Furthermore, each parameter has an average of 16 items varying in form, content, and difficulty level. This ensures that each parameter has a sufficient number of items to measure different aspects of the specific skill. For example, Verbal skills cover both understanding of words and grammar as well as correct use of language. The varying difficulty levels of the tasks also ensure an adequate basis for the score on each parameter, regardless of the test person's general level.

Figure 5.1. Hierarchical representation of Logics parameters and their relation to classical intelligence theory.



Construct validity

Construct validity concerns the agreement between test results and prior theoretical knowledge of the construct being measured. Below, construct validity is documented by studies of convergent validity, where Logics has been compared to well-documented intelligence tests. Also, the internal structure of the test was investigated with item-scale correlations, scale correlations and factor analysis.

Convergent validity (Raven's Progressive Matrices)

The first study consists of a comparison between Logics and Raven's Progressive Matrices (Raven, 1998), which is one of the most recognized test tools in research. Like Logics, Raven's Progressive Matrices is based on Spearman's *g* factor. Although the task types in Logics differ to some extent from those in Raven, which are primarily based on figures, the shared theoretical basis provides an opportunity to compare the overall GMA (General Mental Ability).

The study included 99 people (of which 50 were women) who completed both tests in counterbalanced order. Next, the relationship between the two intelligence quotients was examined with a correlation (Pearson's *r*) shown in Table 5.1. In addition, an independent samples t-test showed that the mean GMA score was not significantly different between orders, $t(97) = 1.24$, $p = .217$.

Table 5.1. Correlation between GMA on Logics and Raven's Progressive Matrices.

Convergent validity	*r*	*p*	N
Cognitive Capacity	.62	< .001	99

When Raven scores were compared to each of the five parameters in Logics – Perception, Analysis, Complexity, Numerical and Verbal – it was found that all parameters correlated significantly with the intelligence quotient in Raven (where Numerical had the highest correlation of 0.61). This indicates a relationship between the level of intelligence and the ability to solve different types of tasks, which is exactly what Spearman's theory of the *g* factor asserts.

Convergent validity (Adaptive Matrigma)

The second study compared Logics to Adaptive Matrigma. Adaptive Matrigma is an adaptive assessment of General Mental Ability relying solely on non-verbal item content (matrices). Within a total of 12 minutes, the test person is to complete a minimum of 12 and a maximum of 40 items with a fixed time limit of 60 seconds per item (Mabon et al., 2017).

3,000+ participants who completed Logics in a recruitment (high-stake) setting were invited to participate in the study. A total of 280 participants signed up for the study, of which 180 (64 %) completed Adaptive Matrigma in a research (low stake) setting. The final sample consisted of 56 % male and 44 % female participants aged 21-65 (M = 43.7, SD = 11.4). A vast majority (98,9 %) stated Danish as their nationality and native language, in which Logics was also completed. The sample comprised 33 % employees, 42 % specialists, and 18 % managers, and a total of 68 % reported a bachelor's degree or higher as their highest completed education (ISCED level 6-8).

Table 5.2 shows the correlations between the overall GMA score from Adaptive Matrigma and each of the scales in Logics. The Spearman correlation is reported for the full sample and the Pearson correlation for the sample excluding outliers. These were identified based on the following criteria:

- Abnormally high number of items completed (40)
- Unexpected low percentage of items answered correctly suggesting guessing, quitting, and/or low effort

In total, 14 outliers were identified that is likely attributable to the low-stake purpose of completing Matrigma. When outliers were removed from the data, the Pearson correlations showed similar magnitudes, and the rank-order of correlations did not change significantly (Cognitive Capacity, Learning Capability, and Accuracy remained the top 3 predictors).

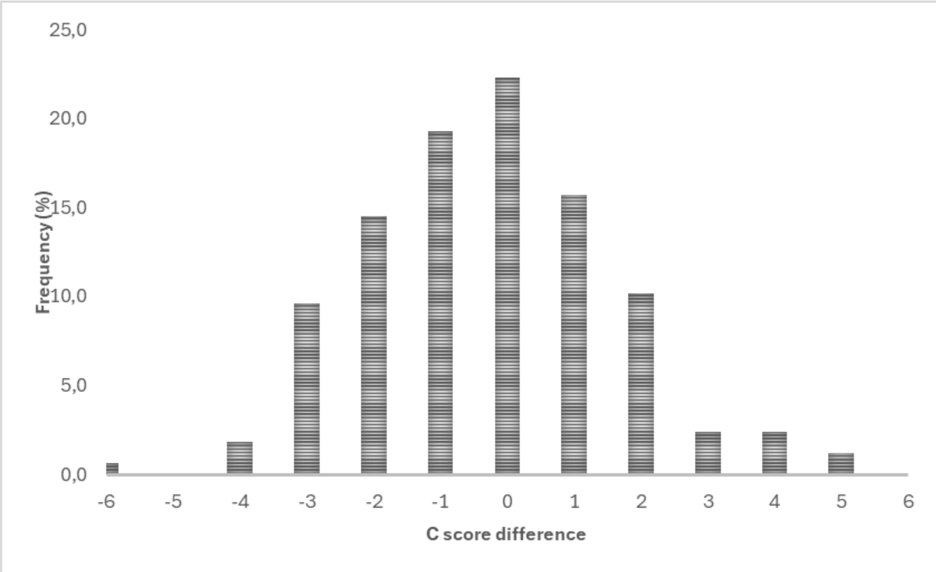
Table 5.2. Correlations between GMA in Adaptive Matrigma and scales in Logics.

Domain	Scale	Pearson*	Spearman
GMA	**Cognitive Capacity**	**.49**	**.51**
	Learning Capability	**.49**	**.52**
Decision Style	Spees	.24	.23
	Accuracy	.50	.55
Logical Capacity	Perception	.45	.46
	Analysis	.41	.46
	Complexity	.43	.46
Skills	Numerical	.37	.36
	Verbal	.28	.28

**Note. N = 166 (14 outliers excluded*).

In addition to the correlations, we also examined score differences at the C score level between GMA as measured in Matrigma and Cognitive Capacity as measured in Logics. The distribution of differences for the sample excluding outliers (N = 166) is shown below in Figure 5.2.

Figure 5.2. C score difference between Adaptive Matrigma and Logics.



For absolute values, 57 % had a C score difference of 0-1 between the assessments, 37 % had a difference of 2-3 C scores, and 6 % had a C score difference of 4 points or higher.

In conclusion, the study shows acceptable convergent validity for the GMA score in Logics and modest score differences between the assessments. Furthermore, the results confirm expected relationships given the difference in methodology between the two assessments:

- Stronger correlations between Adaptive Matrigma and fluid intelligence (Learning Capability and Logical Capacity).
- Weaker correlations between Adaptive Matrigma and crystallized intelligence (Skills).
- Stronger correlations with Adaptive Matrigma for Accuracy compared to Speed.

Item-scale correlations

Next, the construct validity of Logics was examined by correlating scores on individual items with scores on the various parameters (scale scores). Table 5.2 lists the correlations for the first two items belonging to each of the five parameters in Logics (Perception, Analysis, Complexity, Numerical, and Verbal). As can be seen from correlations highlighted in bold, the individual items correlate the strongest with the scale to which they belong. Items P1 and P2 thus correlate the strongest with Perception, whereas items A1 and A2 correlate the strongest with Analysis and so on. Furthermore, all correlations are highly statistically significant. These correlations show that each item measures exactly the parameter for which it was developed. In other words. the item-scale score correlations document the internal coherence and structure of the test.

Table 5.2. Correlations between items and scales in Logics.

Item	Scale				
	Perception	Analysis	Complexity	Numerical	Verbal
P1	**.25**	.09	.11	.09	.08
P2	**.46**	.15	.22	.15	.14
A1	.15	**.35**	.17	.15	.17
A2	.14	**.30**	.09	.11	.10
C1	.11	.07	**.36**	.08	.07
C2	.16	.08	**.48**	.12	.09
N1	.15	.12	.21	**.37**	.11
N2	.08	.08	.11	**.21**	.07
V1	.08	.12	.08	.09	**.30**
V2	.02	.03	.04	.02	**.21**

Scale correlations

In further support of the construct validity for Logics, expected patterns of correlations between scales (parameters) were observed, listed in the correlation matrix in Table 5.3. As expected, all scales showed moderate to strong internal correlations reflecting the notion of the *g* factor. Importantly, GMA correlated almost equally strongly with Speed and Accuracy, showing that the measure of GMA is not inadvertently affected by different strategies (prioritizing speed at the cost of accuracy or vice versa). Unlike GMA, the measure of Learning showed a markedly stronger correlation with Accuracy than Speed and somewhat higher correlations with parameters in Logical capacity compared to Skills (which is expected from how this scale was constructed and the limited learning effect on items within Skills).

Table 5.3. Correlation matrix for Logics scales.

No	Scale	1	2	3	4	5	6	7	8
1	Speed	-							
2	Accuracy	.08	-						
3	Perception	.58	.49	-					
4	Analysis	.64	.55	.51	-				
5	Complexity	.47	.56	.53	.45	-			
6	Numerical	.77	.47	.62	.61	.55	-		
7	Verbal	.70	.52	.53	.67	.49	.66	-	
8	Learning Capability	.20	.76	.56	.64	.60	.35	.40	-
9	**Cognitive Capacity**	**.71**	**.73**	**.74**	**.82**	**.71**	**.86**	**.84**	**.65**

Factor structure

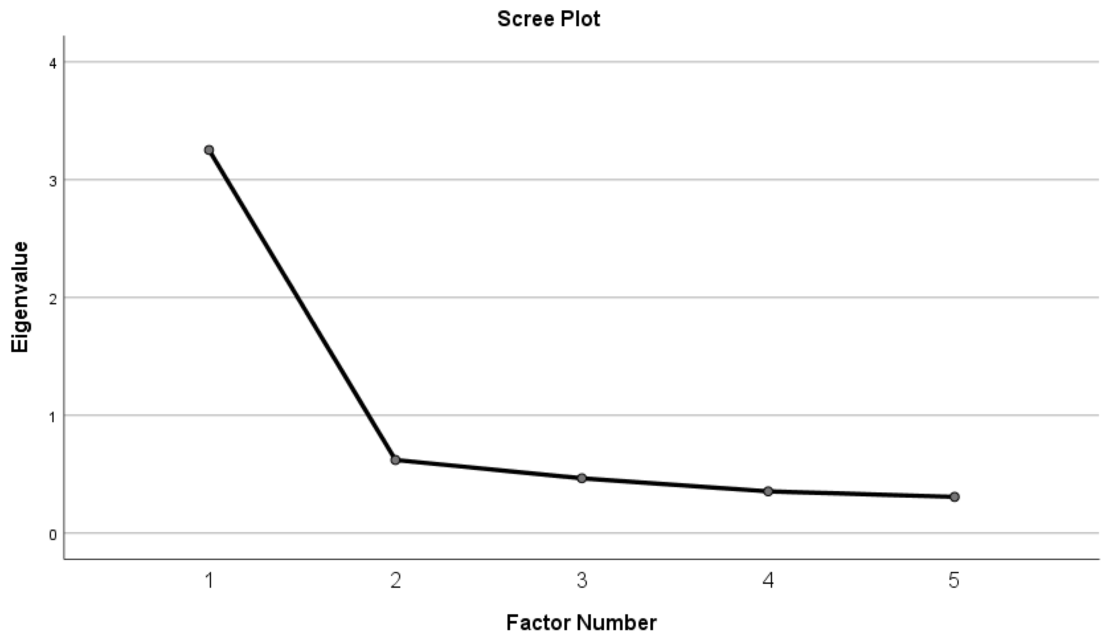
The strong correlation between the overall intelligence quotient and each of the five parameters in Logics is further confirmed by the factor structure of the test, which was examined with an exploratory factor analysis of the five main parameters in Logical capacity and Skills (other scales were omitted due to overlapping or redundant items).

The factor analysis (*Principal Axis Factoring*) showed that one factor should be extracted explaining 65 % of the total variance (supported by the Scree plot below in Figure 5.3) in agreement with the notion of the g factor. Factor loadings for the individual Logics scales are listed below in Table 5.4.

Table 5.4. Factor loadings of Logics scales.

Scale	Factor loading
Perception	.72
Analysis	.75
Complexity	.65
Numerical	.84
Verbal	.80

Figure 5.3. Scree plot from factor analysis of main scales in Logics.



Criterion validity

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

Criterion validity thus involves verifying individual test scores using an external, observable criterion. Studies on both concurrent and predictive criterion validity are conducted on a regular basis in collaboration with customers to examine, for instance, what characterizes existing top performers in a company based on objective performance targets, or compile statistics on how new employees perform in the long term, measured on the basis of relevant KPIs in the organization. Based on this data, we help our customers provide management information, company-specific advice and set up objective ideal profiles based on statistics. This assures and adapts the criterion validity of Logics to the specific organizational context in which it is to be used. As measurements of concurrent and predictive validity are something we do at a customer-specific level to address their context in relation to culture, geography, position, etc., we cannot disclose all results of previous studies due to agreements on confidentiality and secrecy. Instead, we are happy to enter a collaboration to conduct studies of concurrent and predictive criterion validity.

An example of such a study concerned identifying common denominators for salespeople who performed significantly better than their peers. The results of the study showed a trend where complexity in the job and the product to be sold greatly impacted the cognitive requirements of the salesperson. The greater the complexity of the job (product), the higher the level of intelligence among top salespeople compared to their colleagues. Another interesting conclusion was that top salespeople in divisions with less complex products scored lower than their less successful counterparts. Thus, the study indicated that less is sometimes more. A final conclusion of the study was that management in general – regardless of the complexity of the product – required a greater logical capacity than sales. This supports our own studies of differences across job levels, see below.

In support of criterion-related validity more generally, a range of group differences and correlations have been examined for the criteria of job level, education level, age, test purpose, and completion option.

Job level

First, the effect of the four job levels (that make up the norm group for Logics) on the overall score (Cognitive Capacity) was examined. As shown in Table 5.5, the mean score (z) increases with job level, where employees score the lowest and executives the highest. A one-way ANOVA confirmed that differences are statistically significant, $F(4,10163) = 126$, $p < .001$, $\eta^2 = 0.047$. These differences are well in line with the job complexity associated with the different job levels.

Table 5.5. Cognitive Capacity for different job levels.

Job level	Prop. (%)	M	SD
N/A	37.3	0.18	0.98
Employee	27.5	-0.33	0.99
Specialist	26.6	0.09	0.95
Middle manager	6.4	-0.12	1.02
Executive	2.2	0.33	0.86

Educational level

Second, Cognitive Capacity increases with educational level as defined by ISCED (International Standard Classification of Education) shown below in Table 5.6. A one-way ANOVA confirmed that mean scores were significantly different between educational levels, $F(6, 10016) = 121$, $p < .001$, $\eta^2 = .068$.

Table 5.6. Cognitive Capacity for different education levels.

ISCED	Education level	Prop. (%)	M	SD
0-1	Not completed lower secondary	0.05	-0.59	1.68
2	Lower secondary education	1.1	-0.64	1.04
3	High school or vocational education	8.1	-0.48	0.96
4-5	Short-cycle higher education	2.2	-0.69	0.85
6	Bachelor's degree	30.2	-0.17	0.98
7	Master's degree	48.6	0.18	0.97
8	Research degree (PhD. doctorate)	9.6	0.30	0.92

Age

Third, meaningful correlations have been identified between age and scores on the various parameters in Logics. As shown in Table 5.7, Logical capacity – particularly the parameters Perception and Complexity – correlates negatively with age, whereas age does not correlate significantly with neither overall Skills nor GMA. These correlations are consistent with a number of international studies showing that fluid intelligence – in Logics represented by the parameters under Logical capacity – tends to decrease with age, whereas crystallized intelligence (i.e., Skills) remains relatively intact across the life span (Ryan et al.,2000). The result also shows that the intelligence quotient is not affected by age, and that Logics thus does not discriminate inappropriately against different age groups.

Table 5.7. Correlations between age and scales in Logics.

Parameter	Correlation (age)
Logical capacity	**-.07**
Perception	-.08
Analysis	.01
Complexity	-.09
Skills	**.03**
Numerical	.02
Verbal	.04
Cognitive Capacity	**.02**

Test purpose

It has also been investigated how the purpose of completing the test affects results. As shown in Table 5.8, whether the analysis is used for recruitment or personal development entails no statistically significant differences on any scale. In other words, the scales of Logics are independent of whether the test is used in high or low stake situations.

Table 5.8. Statistical comparison of Logics parameters in relation to test purpose.

Domain	Parameter	Recruitment		Development		Comparison	
		M	*SD*	*M*	*SD*	*t*	*p*
Logical capacity	Perception	48.2	18.4	47.7	18.7	0.36	0.72
	Analysis	51.0	16.4	50.7	15.4	0.24	0.81
	Complexity	50.4	19.2	49.3	20.0	0.71	0.48
Skills	Numerical	52.5	20.7	51.3	22.2	0.71	0.48
	Verbal	50.4	17.6	50.1	17.6	0.21	0.83
Cognitive Capacity		**0.06**	**0.93**	**0.03**	**0.93**	**0.42**	**0.68**

Completion option

A final aspect concerns the options of completion. Traditionally, there have been two ways to complete Logics: 1) Screen version and 2) Test link. The screen version was used in a controlled environment, i.e. completed at the company without access to aids other than paper and pen, and the administrator started the test on a computer specifically intended for the purpose. In its form, the test link is completely identical to the screen version, but the test is completed at home, and it is the test subject who clicks the link, starts the test and is accountable for a fair and honest test completion using no other aids than pen and paper.

As shown in Table 5.9, Numerical is the only parameter where the mean score is significantly higher for the test link than the screen version. This may be due to several factors, e.g. that aids can be used in the form of a calculator or external help, or that the test is completed in a familiar environment, and when the test subject feels the most energetic and rested.

Table 5.9. Comparison of completion options for scales in Logics.

Scale	Test link		Screen version		Comparison	
	M	*SD*	*M*	*SD*	*t*	*p*
Perception	48.5	18.4	46.4	19.0	1.46	0.14
Analysis	51.2	16.4	51.2	16.6	-0.04	0.97
Complexity	50.6	19.2	48.5	21.5	1.37	0.17
Numerical	53.0	20.5	45.8	24.4	4.37	< .001
Verbal	50.4	17.6	51.4	18.2	-0.70	0.48
Cognitive Capacity	**0.08**	**0.93**	**-0.04**	**1.03**	**1.70**	**0.09**

We still recommend administering Logics in-house in a controlled setting, though our platform no longer differentiates between the two forms of administration (even if most scales are not affected by completion option).

With the introduction of and democratization of AI, this recommendation is becoming increasingly important. As mentioned in section 3 on administration, scales scores are continuously monitored for drifting of means and higher percentiles (P93), which would be expected in the case of widespread cheating. However, the numbers shown in Figures 5.4 and 5.5 show that this is not a major cause of concern. In addition, standardized mean differences (Cohen's d) between 2022 Q2 and 2025 Q2 were only 0.02 and 0.06 for Numerical and Verbal, respectively.

Figure 5.4. Quarterly monitoring of scale scores on Numerical.

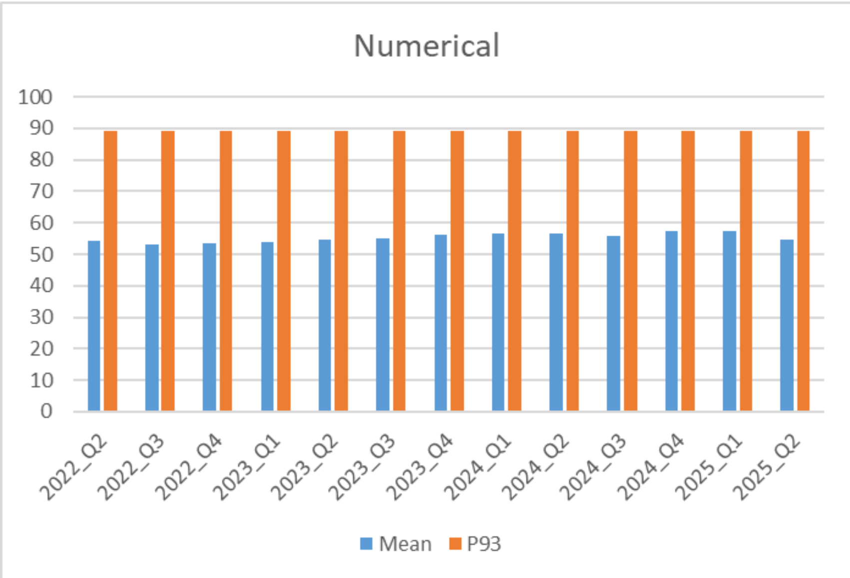
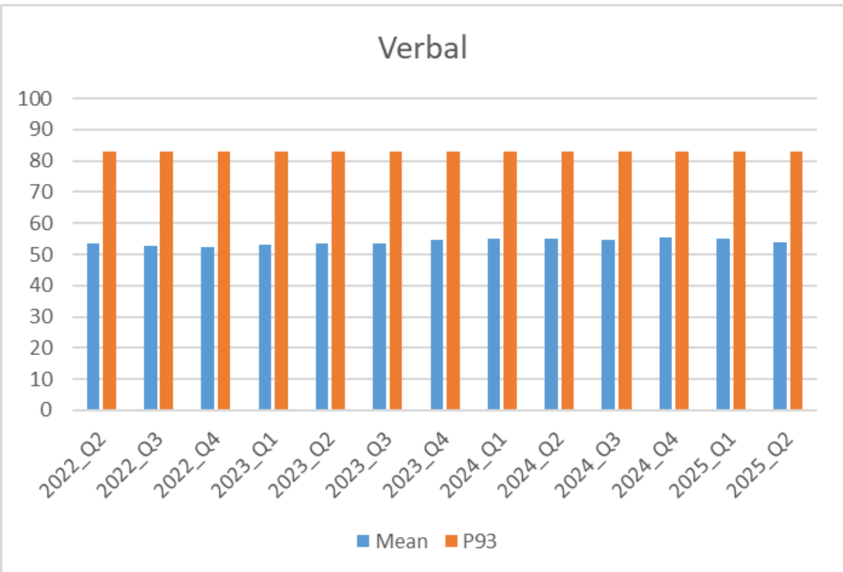


Figure 5.5. Quarterly monitoring of scale scores on Verbal.



6. Reliability

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

Often-used measures of internal consistency are Cronbach's alpha (Cronbach, 1951) and split-half reliability (Spearman-Brown corrected), both of which are listed below in Table 6.1. The reliability estimates are based on data collected between November 2019 and March 2020. In accordance with EFPA guidelines, the group of test subjects is composed based on the same criteria as the norm group in order to match the intended target group. The following studies are thus based on a group of 3,288 test subjects who are office workers at the white-collar level and above, aged 20-65 years (M = 38.4, SD = 10.5). The sample comprised an equal number of males and females and 17 % managers (middle managers or executives).

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

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

The SEMs are reported with z-scores for Cognitive Capacity and Learning Capability and raw scores for the remaining scales.

Where SD represents the standard deviation, and r refers to the reliability of the scale in question.

Table 6.1. Reliability estimates and standard error of measurement (SEM) for scales in Logics.

Scale	Alpha	Split-half	SEM	SEM (C score)
Cognitive Capacity	**.89**	**.87**	**0.33**	**0.66**
Learning Capability	**.66**	**.73**	**0.58**	**1.17**
Logical Capacity	**.76**	**.74**	**2.69**	**0.98**
Perception	.50	.56	1.30	1.14
Analysis	.62	.69	1.85	1.23
Complexity	.54	.60	1.31	1.36
Skills	**.83**	**.85**	**2.53**	**0.82**
Numerical	.79	.84	1.71	0.92
Verbal	.68	.74	1.79	1.13

The result shows a total coefficient alpha and split-half reliability for Cognitive Capacity of .89 and .87, respectively, which is close to the excellent mark reported by EFPA (2013). However, please note that for individual scales at the lowest level, reliability coefficients are not sufficiently high to warrant basing recruitment decisions on single scores only. Rather, decisions should be based on composite scales, ideally the total Cognitive Capacity score (or total Logical capacity and Skills). For these scales, the Standard Error of Measurement (SEM) is also low. i.e., less than 1 C score point.

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. Logics is a normative test, which means that the test result is compared to a relevant norm group.

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).

As opposed to power or adaptive tests (where the item difficulty increases throughout the test or adjusts based on the test person's ability), Logics is a classical speed test. This means that items appear in a fixed, randomized order, and most items are fairly easy to solve (with most item difficulties in the .60-.90 range). Rather than increasing or adjusting item difficulty, the test person is challenged with having many tasks to complete within a set time limit.

Score calculation

First, the number of correct answers is summed and divided by the total number of items to form a raw score for each scale (representing the total proportion of correct answers). 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, the difficulty of items and hence scores on any cognitive assessment might change because of increased familiarity with test formats, item stimuli or even the Flynn effect (although this effect appears to have decreased over the past years). Therefore, with respect to assessments, it is highly important to update norm groups on a regular basis and make sure that test persons are evaluated with a norm group representing the current state of affairs, 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.

The current Logics norm was updated in 2025 based on data collected in a high-stake setting (selection and development) from 2020-2024.

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).

The global norm group for Logics consists of 10,168 people who were selected through stratified randomization from a total of 29,277 people aged 18-70 who completed the test in a high-stake setting. Statistical analyses confirmed that the norm group does not represent a biased sample, as score differences between included and excluded samples were only small or negligible across scales (Cohen's d ranging from 0.00-0.13 with an average of 0.04).

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. To construct a proper global norm, the sample was stratified for gender at the nationality level, hence making each nationality contributing 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. Too few people stated "other" as gender to be represented as an individual group. Therefore, their preferred pronouns (he/him vs. she/her) were used to include them in the binary gender groups.

Although the final age distribution was slightly skewed to the left (median of 34), statistical analyses showed small correlations between age and any of the Logics scales with absolute values ranging from .05 to .16 averaging .11. 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 and job levels, the sample was not further stratified for any of these demographic variables.

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

Table 7.2. Demographic composition of the global norm group for Logics.

Global norm: Logics	
Last updated	2025
Data collection	2020-2024
Sample size	10,168
Composition	
Purpose(s)	Selection: 91.5 % Development: 8.5 %
Gender	Male: 50.0 % Female: 50.0%
Age	18-70 (M = 35.0, SD = 8.71)
Nationalities	126 (max. = 2.5 %)
Education level (%)	
N/A	1.4
Primary & lower secondary school	1.1
Vocational	3.6
High school	4.5
Short-cycle higher education	2.2
Bachelor's degree	29.8
Master's degree	47.9
Research degree (PhD, doctorate)	9.5
Job level (%)	
N/A	37.3
Employee	27.5
Middle manager	6.4
Specialist	26.6
Executive	2.2

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.

Simulations of expected AI were conducted at different selection rates for gender (male/female) and age (below/above 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. Generally, combining methods decreases adverse impact but not if the combined methods show impact on the same group. Hence, it is important to be mindful of potential adverse impact at a scale level, even with the recommended application of assessments.

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 (below/above 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.

For gender, no adverse impact is expected for any scale even for strict selection ratios, as AI ratios ranged from 0.80 to 0.98. For age, careful consideration should be paid to a possible adverse impact favoring younger individuals when applying strict or moderate selection ratios on Speed, Perception, Complexity, Numerical, and Cognitive Capacity. If these scales are used as the basis for selection (either individually or in combination), careful consideration should be given to avoid fixed and strict (in some cases moderate) selection ratios and combine scores on this scale with other criteria to balance out the level of Adverse Impact, thus preventing any indirect discrimination on age.

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

Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
Cognitive Capacity	**2.5**	**0.10**	**0.83**	**0.87**	**0.94**
Learning Capability	0.4	0.07	0.88	0.93	0.97
Speed	1.1	0.06	0.90	0.92	0.96
Accuracy	0.9	0.07	0.86	0.93	0.96
Perception	1.3	0.06	0.88	0.93	0.95
Analysis	1.4	0.08	0.87	0.90	0.94
Complexity	-0.2	0.01	0.98	0.99	0.97
Numerical	3.7	0.17	0.80	0.84	0.89
Verbal	-0.2	0.01	0.98	0.99	1.00

Table 7.4. Adverse Impact simulations for age (below/above 40) at different selection ratios (SR).

Scale	Dif.	d	Strict SR	Moderate SR	Lenient SR
Cognitive Capacity	**6.0**	**0.24**	**0.74**	**0.77**	**0.82**
Learning Capability	0.6	0.09	0.94	0.91	0.92
Speed	4.2	0.23	0.73	0.81	0.83
Accuracy	1.6	0.13	0.83	0.87	0.94
Perception	4.9	0.25	0.69	0.77	0.81
Analysis	1.9	0.11	0.89	0.89	0.93
Complexity	5.2	0.26	0.66	0.74	0.82
Numerical	6.5	0.29	0.62	0.71	0.80
Verbal	2.7	0.15	0.82	0.86	0.90

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

8. Translations and Adaptations

The following section outlines the process for translations and adaptation of items.

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 proof-reads the translations 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 the 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, key words or specific numbers must not be changed to prevent numerical or verbal tasks from becoming more difficult or easier in the target language.
- **Source versions**: Ideally, the original version (typically 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.

- For Logics specifically, all items are localized such that names, currencies, measurement units, proverbs, etc. match the target population (for English, separate versions are made for UK and US). Such changes may not alter the content or difficulty of the item; however, the context can be adjusted if needed (e.g., changing which item is bought for X amount of money in the local currency).

References

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19*(3), 3127-3131.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

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.

Flanagan, D. P. & Dixon, S. G. (2014). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In *Encyclopedia of Special Education* (eds C.R. Reynolds, K.J. Vannest and E. Fletcher-Janzen).

Mabon, H., Niemelä, F., Sjöberg, A., & Sjöberg, S. (2017). *Adaptive Matrigma – Technical manual*. Assessio International AB.

Raven, J. C. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Ryan, J. J., Sattler, J. M., & Lopez, S. J. (2000). Age effects on Wechsler adult intelligence scale-III subtests. *Archives of Clinical Neuropsychology*, *15*(4), 311-317.

Schmidt, F. L., Oh, I. S., & Shaffer, J. A. (2016). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings. *Fox School of Business Research Paper*.